Volume 101
Number 4
November 2009

Published quarterly
by the
American Psychological
Association

ISSN 0022-0663

Journal of Educational Psychology

www.apa.org/journals/edu

The main purpose of the *Journal of Educational Psychology*® is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Change of Address: Send change of address notice and a recent mailing label to the attention of Subscriptions Department, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee periodicals forwarding postage.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; call 202-336-5600 or 800-374-2721; or visit www.apa.org/journals/subscriptions.html.

Microform Editions: For information regarding microform editions, write to University Microfilms, Ann Arbor, MI 48106.

Manuscripts: Submit manuscripts electronically through the Manuscript Submissions Portal found at www.apa.org/journals/edu according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Editor, Art Graesser, Journal of Educational Psychology, 202 Psychology Building University of Memphis, Memphis, TN 38152-3230. The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

Electronic Access: APA members who subscribe to this journal have automatic access to all issues of the journal in the PsycARTICLES® full-text database. See http://my.apa.org/access.html.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

APA Journal Staff: Susan J. A. Harris, *Senior Director, Journals Program;* Skip Maier, *Director, Journal Services;* Paige W. Jackson, *Director, Editorial Services;* Clark Munsell, *Account Manager;* Annie Hill, *Editorial Supervisor;* Amy O'Keefe, *Lead Editor;* Jodi Ashcraft, *Director, Advertising Sales and Exhibits.*

The **Journal of Educational Psychology**® (ISSN 0022-0663) is published quarterly (February, May, August, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2010 rates follow: *Nonmember Individual:* $167 Domestic, $193 Foreign, $204 Air Mail. *Institutional:* $525 Domestic, $569 Foreign, $582 Air Mail. *APA Member:* $78. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to the *Journal of Educational Psychology,* American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

Effective with the 1986 volume, this journal is printed on acid-free paper.

Journal of

Educational Psychology®

November 2009

Volume 101
Number 4

www.apa.org/journals/edu

Articles

(*Contents continue*)

Other

New Editors Appointed, 2011–2016

The Publications and Communications Board of the American Psychological Association announces the appointment of 3 new editors for 6-year terms beginning in 2011. As of January 1, 2010, manuscripts should be directed as follows:

- *Developmental Psychology* (http://www.apa.org/journals/dev), **Jacquelynne S. Eccles, PhD,** Department of Psychology, University of Michigan, Ann Arbor, MI 48109
- *Journal of Consulting and Clinical Psychology* (http://www.apa.org/journals/ccp), **Arthur M. Nezu, PhD,** Department of Psychology, Drexel University, Philadelphia, PA 19102
- *Psychological Review* (http://www.apa.org/journals/rev), **John R. Anderson, PhD,** Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213

Electronic manuscript submission: As of January 1, 2010, manuscripts should be submitted electronically to the new editors via the journal's Manuscript Submission Portal (see the website listed above with each journal title).

Manuscript submission patterns make the precise date of completion of the 2010 volumes uncertain. Current editors, Cynthia García Coll, PhD, Annette M. La Greca, PhD, and Keith Rayner, PhD, will receive and consider new manuscripts through December 31, 2009. Should 2010 volumes be completed before that date, manuscripts will be redirected to the new editors for consideration in 2011 volumes.

Journal of Educational Psychology
2009, Vol. 101, No. 4, 765–778

0022-0663/09/$12.00 DOI: 10.1037/a0015956

Predicting Reading Comprehension in Early Elementary School: The Independent Contributions of Oral Language and Decoding Skills

Panayiota Kendeou
McGill University

Paul van den Broek
Leiden University

Mary Jane White
University of Memphis

Julie S. Lynch
Saginaw Valley State University

The authors examined the development of oral language and decoding skills from preschool to early elementary school and their relation to beginning reading comprehension using a cross-sequential design. Four- and 6-year-old children were tested on oral language and decoding skills and were retested 2 years later. In all age groups, oral language and decoding skills formed distinct clusters. The 2 clusters were related to each other in preschool, but this relation became weaker in kindergarten and 2nd grade. Structural equation modeling showed that both sets of skills in 2nd grade independently predicted a child's reading comprehension. These findings confirm and extend the view that the 2 clusters of skills develop early in a child's life and contribute to reading comprehension activities in early elementary school, with each cluster making a considerable, unique contribution.

Keywords: reading comprehension, oral language, decoding, development

Despite intensive instruction, many children in early elementary school fail to reach functional levels of literacy. As a result, there has been considerable research on early intervention programs for fostering decoding skills—skills that support decoding, such as phonological awareness and letter and word identification (see Ehri, Nunes, Stahl, & Willows, 2001; Papadopoulos, Das, Parrila, & Kirby, 2003; Snowling & Hulme, 2005; Storch & Whitehurst, 2002, for reviews). There also is growing recognition that the development of these skills should be complemented by fostering oral language skills[1]—skills that support comprehension, such as receptive vocabulary (i.e., understanding of spoken words) and narrative comprehension (e.g., Lonigan, Burgess, & Anthony, 2000; Pressley et al., 2001; see also the influential "simple view of reading," Gough & Tunmer, 1986; Hoover & Gough, 1990; Tunmer & Hoover, 1992). However, the findings with respect to the relative contribution of oral language skills in early reading comprehension have been contradictory.

On the one hand, the results of some studies suggest that oral language skills are not central to early reading comprehension and that they become fully operative only when the child has acquired decoding skills (Bryant, McLean, & Bradley, 1990; Speece, Roth, Cooper, & de la Paz, 1999; Vellutino, Tunmer, Jaccard, & Chen, 2007). On the other hand, results of other studies suggest the opposite, highlighting the importance of such skills in early reading comprehension (Bishop & Adams, 1990; Catts, Fey, Zhang, & Tomblin, 1999; Paris & Paris, 2003).

The lack of consensus in the literature may be the result of several limitations. One issue concerns the different ways that oral language skills and reading comprehension skills have been conceptualized and measured. The far-reaching implications of such variation, in both predictor and outcome measures, have been discussed recently with respect to several widely used tests of reading comprehension (Cutting & Scarborough, 2006; Fletcher, 2006; Keenan & Betjemann, 2006; Magliano, Millis, Ozuru, & McNamara, 2007; Ozuru, Rowe, O'Reilly, & McNamara, 2008; VanderVeen et al., 2007). Indeed, there is direct evidence that commonly used tests of reading comprehension are not tapping into the same cognitive processes (Kendeou & Papadopoulos,

Panayiota Kendeou, Department of Educational and Counselling Psychology, McGill University, Montreal, Quebec, Canada; Paul van den Broek, Department of Pedagogy, Leiden University, Leiden, the Netherlands; Mary Jane White, Department of Psychology, University of Memphis; Julie S. Lynch, Department of Psychology, Saginaw Valley State University.

This project was supported by grants from the Center for the Improvement of Early Reading Achievement (CIERA) at the University of Michigan School of Education, the Center for Cognitive Sciences at the University of Minnesota, and the U.S. National Institute of Child Health and Human Development (Grant No. HD-07151). Writing of this article was supported by a Fonds Quebecois de la Recherche sur la Societe et la Culture (FQRSC) Grant No. 2009-NP-125707 to Panayiota Kendeou and a Golestan and Lorentz fellowship from the Netherlands Institute for Advanced Study to Paul van den Broek. We are grateful to Danielle S. McNamara for her comments on the article.

Correspondence concerning this article should be addressed to Panayiota Kendeou, Department of Educational and Counselling Psychology, 3700 McTavish Avenue, Montreal, Quebec H3A 1Y2, Canada. E-mail: panayiota.kendeou@mcgill.ca

[1] There are many different terms in the literature for oral language skills (e.g., oral comprehension, language comprehension). We used the term *oral language skills* in the present study because we included receptive vocabulary in addition to narrative comprehension when estimating the construct.

2009; RAND Reading Study Group, 2002), and as a result, there is a renewed interest in the development of theory-based assessments of comprehension (Francis et al., 2006; Francis, Fletcher, Catts, & Tomblin, 2005).

Another issue concerns the age range on which existing studies have focused. With only a few exceptions, the majority of the studies assessing oral language and decoding-related skills and their usefulness as predictors of reading comprehension have investigated children in kindergarten through second grade (for a review, see Storch & Whitehurst, 2002). Many of these skills, however, develop well before kindergarten, so research focusing on younger children is needed to provide a full understanding of the developmental trajectories of certain skills. A final issue is that the majority of these studies have been cross sectional. Longitudinal investigations would more accurately reflect developmental patterns.

In the present study, we explored the relative contributions and interrelations of oral language and decoding skills in reading comprehension, while addressing the aforementioned issues by (a) using a theoretical framework to guide our assessments, (b) starting with preschool children, and (c) adopting a cross-sequential longitudinal design. Specifically, we conceptualized and measured oral language and reading comprehension skills within the framework of the causal network theory (CNT; see Goldman & Varnhagen, 1986; Graesser & Clark, 1985; Stein & Glenn, 1979; Trabasso, Secco, & van den Broek, 1984) and followed longitudinally two cohorts of children in preschool and kindergarten over the course of 2 years (i.e., a cross-sequential longitudinal design). This design allowed for longitudinal yet time-efficient data collection. More important, it allowed us to compare consecutive age groups both across the same materials (the cross-sectional component) and within the same individuals across time (the longitudinal component).

Assessing Comprehension Within the CNT Framework

Although researchers and educators use the term *comprehension* in different ways, there is considerable agreement that central to comprehension in the context of reading is the construction of a coherent mental representation of the text. In this representation, the reader successfully connects statements and ideas from the text. This mental representation is based on both the text itself and the readers' background knowledge (e.g., Applebee, 1978; Gernsbacher, 1990; Graesser, Singer, & Trabasso, 1994; Kintsch & van Dijk, 1978; Mandler & Johnson, 1977; Oakhill & Cain, 2007; Pearson & Hamm, 2005; Stein & Glenn, 1979; Trabasso et al., 1984; van den Broek, 1994). Thus, comprehension depends on knowledge that cannot always be found in a single word or sentence (Whitehurst & Lonigan, 1998) or even in the text proper. In addition, comprehension is not a unitary phenomenon but rather a family of skills that develop simultaneously (Cutting & Scarborough, 2006; Duke, 2005; van den Broek et al., 2005; Vellutino et al., 2007). This family of skills includes higher order processes—such as inference generation and reasoning—that enable readers to identify meaningful relations among text elements and between text elements and background knowledge.

Events in a text can be related in many ways, but one particularly important type of connection is causal (Graesser et al., 1994; van den Broek, 1997). When readers engage in making inferences and generate different types of causal connections to interconnect the events of the narratives they read, they form a mental network representation of the narrative. This mental network is based on the causal structure of the narrative itself. Consider, for example, the excerpt from the story *The Cat's Purr* in Figure 1. In the accompanying network, each clause in the story is represented in the network as a circle with the corresponding number. The arrows between the circles represent the causal connections between the clauses that the reader may identify. For instance, when reading the story, the reader may make the connection that the fact that Rat wanted to copy Cat (Sentence 4) causes several actions such as Rat's building a house like the house that Cat built (Sentence 5) and planting a tree (Sentence 7).

Such networks have been found to capture important aspects about adults' comprehension of stories (e.g., Goldman & Varnhagen, 1986; O'Brien & Myers, 1987; Trabasso & van den Broek, 1985). Adult readers are more likely to remember events with many causal connections than events with few connections. For example, readers are more likely to remember that Rat liked to copy Cat (Sentence 4; a statement that has five connections) than that Cat let Rat play a tune (Sentence 13; a statement that has one connection). In addition to better remembering events with many connections, readers tend to include these events in summaries more often and rate them as more important than events with only a few connections (van den Broek, 1986; van den Broek, Lorch, & Thurlow, 1996). In addition, when readers are reminded of one event in the text, they remember related events more quickly than unrelated events, even when the latter are closer together in the original text (e.g., O'Brien & Myers. 1987; van den Broek & Lorch, 1993). Finally, when readers are asked questions about why

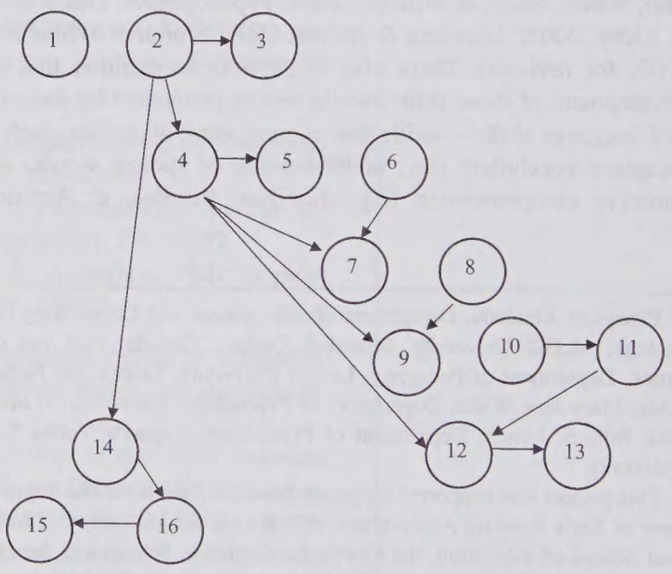


Figure 1. Excerpt from *The Cat's Purr,* the aural story used in Year 1: 1→ Once upon a time, 2→ Cat and Rat were best friends. 3→ They lived in houses right next to each other. 4→ Rat liked to copy Cat. 5→ Rat built a house that was just like Cat's. 6→ Cat planted a tree by his house. 7→ Rat planted one too. 8→ Cat made a straw mat for his house, 9→ and Rat made one too. 10→ Once, Cat made a flute 11→ and played sweet tunes. 12→ "Let me play a tune, too," said Rat. 13→ So, Cat let Rat play a tune. 14→ Cat and Rat also worked together in their vegetable garden. 15→ They planned to have a big party for their friends 16→ when all of the vegetables were ready.

an event happened, they respond with answers that are based on the causal connections in the text (Graesser & Clark, 1985; Trabasso, van den Broek, & Liu, 1988). For example, when asked why Rat built a house that was just like Cat's (Sentence 5), readers are very likely to answer that this was because Rat liked to copy Cat (Sentence 4).

With age, children's sensitivity to the causal structure of narratives as well as their ability to generate causal inferences necessary to connect ideas in the text and to form a coherent text representation develop (Ackerman, 1988; Casteel, 1993; Trabasso & Nickels, 1992; van den Broek, 1988, 1989a, 1990). At an early age, nonstructural properties of events, including superficial ones such as the event's vividness, influence young children's attention more so than do causal structure properties. At later ages, the role of structural properties increases and that of nonstructural factors decreases (Bourg, Bauer, & van den Broek, 1997; van den Broek, 1997). Young children also have a tendency to focus much of their attention on observable, concrete actions rather than on internal causes such as a character's goals. As comprehension skills improve, children's understanding of motives attributable to a character's action, event, or state improves (Goldman & Varnhagen, 1986; Mandler, 1984; Stein & Glenn, 1979; van den Broek, 1989a). Finally, children's tendency to limit their connection building to events *within* each episode gradually diminishes and building possible connections *between* episodes increases. These latter connections are most likely to be related to the overall theme or message of the text and, hence, are central to obtaining a complete picture of the meaning of the narrative as a whole (Brown & Smiley, 1978; Goldman & Varnhagen, 1986; Trabasso & Nickels, 1992; Williams, 1993).

Observations about the development of children's comprehension have primarily involved elementary school children. Research on children's comprehension processes before elementary school is more limited. Because preschool children cannot read, assessment of their comprehension necessarily involves a nonreading context in which stories are presented in media other than text. For example, stories can be presented using pictures (e.g., Paris & Paris, 2003), aurally (e.g., Stein & Glenn, 1979; Trabasso et al., 1984), or via television (e.g., Lorch, Bellack, & Augsbach, 1987; van den Broek et al., 1996). Studies on event comprehension have revealed several developmental patterns from infancy to adulthood and have shown that even young children are sensitive to the relations that exist between different events and particularly causal relations (Mandler & Johnson, 1977; Stein & Glenn, 1979). In fact, children as young as 2 years old can identify different types of causal relations in event sequences (Bauer, 1996, 1997; Wenner & Bauer, 2001). These results suggest that although there are clear developmental differences, even very young children, when comprehending events, use cognitive processes that are similar to those older children and adults use: They identify connections and generate inferences (Lorch & Sanchez, 1997; Mares, 2006; Trabasso & Nickels, 1992; van den Broek et al., 1996). Moreover, these results also suggest that similar processes contribute to comprehension of narratives across different media (also see Kendeou, van den Broek, White, & Lynch, 2007; Lynch et al., 2008; van den Broek et al., 2005). Thus, using different media provides a unique opportunity to assess comprehension independently of decoding skills and before children begin formal instruction.

The Present Study

The fact that coherence-focused comprehension skills already are present and developing in preschool children raises important questions about the developmental trajectory of oral language skills in individual children. One set of questions concerns whether oral language and decoding skills develop according to independent and unique tracks or whether they are interdependent. The results of several recent studies have shown that within an age group, the correlations between decoding and oral language skills tend to be low, suggesting that at least within age groups, there is no direct relation between the two sets of skills at these ages (Cain, Oakhill, & Bryant, 2004; Catts et al., 1999; Gough & Tunmer, 1986; Kendeou, Savage, & van den Broek, in press; Paris & Paris, 2003; Savage, 2006). There is also evidence, though, that oral language and decoding skills are highly related, especially in early years (Storch & Whitehurst, 2002). Another set of questions concerns the stability and predictive power of oral language skills as a child develops: Are individual differences relatively stable across time, and, hence, are skills at an early age predictive of those at a later age? Moreover, are oral language skills predictive of reading comprehension at a later age?

To answer these questions, we examined oral language and decoding skills of two cohorts of children, 4- and 6-year-olds, and retested them when they were 6- and 8-year-olds, respectively. We have taken three steps in our analyses. First, we examined the relation between the development of children's oral language skills and decoding skills from preschool to early elementary school. To do so, we used structural equation modeling to test two models (one for each cohort) derived from the literature (Storch & Whitehurst, 2002). In these models, oral language and decoding skills are connected during preschool (age 4) and kindergarten (age 6). Because oral language skills precede the development of decoding skills, and therefore, influence their development, the link in the structural equation model flows from oral language skills to decoding skills. Second, in these models, we examined the stability and predictive power of children's oral language skills, namely whether oral language skills at an early age predicted those at a later age and whether oral language skills were predictive of reading comprehension at a later age. Third, we tested a number of alternative, theoretically driven models to provide empirical support for the appropriateness of our hypothetical conceptual models as well as the narrative comprehension measures we used to assess oral language skills.

Method

Participants

Two hundred ninety-seven children participated at two test points, 2 years apart. Complete data were collected from 113 children in the 4- to-6-year-old cohort (mean age at the first test point = 4 years, 6 months; range = from 4 years, 0 months to 4 years, 11 months) and 108 children in the 6- to-8-year-old cohort (mean age at the first time point = 6 years, 4 months; range = from 6 years, 0 months to 6 years, 11 months). At both test points, the 6-year-old children were in or had recently completed kindergarten. At the second test point, the 8-year-old children were in or had recently completed second grade. All participants were from a

large city in the upper Midwest, and their parents traditionally collaborated in research projects at the university at which the study was conducted. The large majority (96%) were White. Consequently, this sample was very homogeneous.

Materials

Oral Language Skills Measures

At each time point, two narratives were used as part of the narrative comprehension assessment (listening and television) and the Peabody Picture Vocabulary Test–III (PPVT–III; Dunn & Dunn, 1997) was used to assess vocabulary.

Listening comprehension. For the listening comprehension task, one narrative was presented aurally (on an audiotape) to the children at each time point. At Time 1, the story, *The Cat's Purr*, was a 7-min narrative. At Time 2, the story, *The Rabbit and the Moon*, was a 10-min narrative. Both narratives were based on American folk tales. Simple, line-drawn pictures were made to accompany each story. The pictures alone did not convey any major points from the plots. Both stories had a standard but complex structure in which the protagonists made several attempts to achieve their desired goals.

Television comprehension. For the television comprehension task, one narrative was presented audiovisually (on a 26-inch [66.04-cm] color television) to the children at each time point. At Time 1, the audiovisual story, *Autumn Leaves*, was a 12-min episode from an American children's television series, *The Rugrats*. At Time 2, the audiovisual story, *Granny's New Glasses*, was an 18-min episode from an Australian children's television series, *The Adventures of Blinky Bill*. Like the stories in the listening comprehension tests, both stories had a standard but complex structure in which the protagonists made several attempts to achieve their desired goals.

Vocabulary. The PPVT–III (Dunn & Dunn, 1997) was administered as a standardized measure of receptive vocabulary for standard English at both time points. The PPVT–III was selected because it allows students who are nonverbal to participate by pointing to a response. Words are orally given to the respondent by the examiner. The respondent points to a picture that best corresponds to the word. The total raw score is obtained by subtracting the number of errors from the numerical value of the ceiling item (highest word correctly identified). For the ages included in this study, internal consistency, as measured with coefficient alpha, ranged from .93 to .95. Split-half coefficients ranged from .86 to .95.

Decoding Skills Measures

Letter and word identification subtests from Woodcock Reading Mastery Test–Revised (Woodcock, 1987), and the Onset Recognition Fluency (OnRF) measure from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002a, 2002b) were used to assess decoding skills. All measures were administered at Time 1. At Time 2, only word identification was administered as ceiling effects were expected for the other two measures.

Letter identification. The Letter Identification Subtest from the Woodcock Reading Mastery Test–Revised (Woodcock, 1987) was administered. This test was selected because it provides information about a child's ability to identify different letters. In the test, the letters are presented to subjects, and each subject is asked to verbally identify the letter within 5 s. The test begins at an age-appropriate item (basal level) and ends when the child answers six or more consecutive items incorrectly or when the last page of the test has been administered. The total raw score is obtained by subtracting the number of errors from the numerical value of the ceiling item (the last letter or word correctly identified). For the ages included in this study, internal consistency, as measured with split-half coefficients, ranged from .84 to .94.

Word identification. The Word Identification Subtest from the Woodcock Reading Mastery Test–Revised (Woodcock, 1987) was administered. This test was selected because it provides information about a child's ability to read different words. In the test, the words are presented to subjects, and each subject is asked to verbally identify the word within 5 s. It is not assumed that the subject knows the meaning of any word that is correctly identified. The test begins at an age-appropriate item (basal level) and ends when the child answers six or more consecutive items incorrectly or when the last page of the test has been administered. The total raw score is obtained by subtracting the number of errors from the numerical value of the ceiling item (the last letter or word correctly identified). For the ages included in this study, internal consistency, as measured with split-half coefficients, ranged from .97 to .98.

Phonological awareness. The Onset Recognition Fluency (OnRF) measure from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002a, 2002b) was used to assess children's phonological awareness. The examiner presents four pictures to the child, names each picture, and then asks the child to identify (i.e., point to or say) the picture that begins with the sound produced orally by the examiner. For example, the examiner says, "This is sink, cat, gloves, and hat. Which picture begins with /s/?" and the student points to the correct picture. The child is also asked to orally produce the beginning sound for an orally presented word that matches one of the given pictures. The examiner calculates the total number of initial sounds produced correctly in a minute. For the ages included in this study, alternate-form reliability of the OnRF measure is .91.

Reading Comprehension Measure

A written narrative was used to assess children's reading comprehension at Time 2 (8-year-olds only). The story, *The Barber's Wife*, was based on an Indian folk tale. The story had age-appropriate vocabulary and had a standard but complex structure in which the protagonists made several attempts to achieve their desired goal.

Procedure

Time 1

The children were tested individually in a single session by one of three female experimenters. The entire session was videotaped, and the comprehension tests were also audiotaped. After the children had been given a period of time to become comfortable with the experimental setting; they completed the phonological aware-

ness and vocabulary assessments. The children then listened to the aural story, *The Cat's Purr*. The children were instructed to listen closely so they could answer questions after the story was over. While listening to the story, the children had the pictures that accompanied the story available. Immediately after the story was completed, the experimenters asked the children to "tell everything you remember about the story from the beginning." If a child did not recall any narrative events spontaneously, the experimenter asked a more specific question, "What happened at the beginning of the story?" The children were prompted to continue to recall the story (i.e., "What else do you remember?"), until they indicated that they could not recall anything else. Because young children often have difficulty spontaneously reporting their memory for narratives, slightly more specific questions were then asked. For each episode in the narrative (determined by procedures described by Mandler & Johnson, 1977; Stein & Glenn 1979) that a child remembered spontaneously, the experimenter asked a question in the following form: "You remembered X. What happened before X?" and "What happened after X?" These questions were asked only if the child did not recall the episodes that were the focus of the questions.

After a short break, the children completed the letter and word identification tests. The children then viewed the audiovisual narrative, *Autumn Leaves*. The procedure for assessing comprehension of the audiovisual narrative was identical to that of the aural narrative.

Time 2

The children were tested individually in a single session by one of three experimenters (two women and one man). All procedures were identical to those at Time 1, except for the exclusion of the phonological awareness and letter identification tests and the inclusion of the reading comprehension task for the 8-year-old children. The reading comprehension assessment took place after the children completed the word identification task. The procedure for assessing reading comprehension was parallel to that of comprehension of the aural and audiovisual narratives. To allow for longitudinal comparisons, we kept the order of the different tests and stories identical across both time points.

Coding

Prior to data collection in each year, three researchers analyzed the narratives and parsed them into individual events (generally subject–verb phrases). At Time 1, the aurally presented story had 167 events and the audiovisual narrative had 231 events. At Time 2, the aurally presented story had 158 events, the audiovisual narrative had 357 events, and the written text had 99 events. We determined the causal structure of each narrative by identifying the causal relations between all events in the story according to principles of causality (Mackie, 1980; Trabasso, van den Broek, & Suh, 1989).

The selection of stories ensured that they did not differ greatly in length across Times 1 and 2. Furthermore, the causal analysis of all stories showed that there were approximately an equal proportion of highly connected events (i.e., events with four or more connections) in each one. Specifically, the aurally presented story had 31% highly connected events in Time 1 and 30% in Time 2. The audiovisual narrative had 40% highly connected events in Time 1 and 38% in Time 2. The written text had 32% highly connected events.

The children's responses during the free recall after each story presentation (aural, audiovisual, and written) were transcribed verbatim from the videotapes and audiotapes of the experimental sessions. The children's responses were parsed into events, analogously to the parsing of the original narratives. Each recalled event was coded according to a gist criterion to the event that it most closely matched in the corresponding narrative. Recalled events that did not match an event in the story were coded separately and were not included in the following analyses. These events ranged between 11% and 15% of total events recalled across stories and age groups.

Two raters coded the transcripts. For results at Time 1, 20% of the transcripts were coded by both raters to establish and practice the coding scheme. An additional 20% of the transcripts were coded by both raters to determine interrater reliability. Interrater agreement was .62, $p < .01$ (calculated on the basis of each recalled events as an agreement or disagreement in coding). The same coding procedures were followed for the results at Time 2. Here, interrater agreement also was .62, $p < .01$.

For each narrative, a measure of children's sensitivity to the causal structure was calculated. This measure consisted of the total number of unique, highly connected story events (i.e., events with four or more causal connections to other events in the story) the children included in their free recall protocol (Graesser & Clark, 1985; Trabasso et al., 1989; van den Broek et al., 1996; van den Broek, 1989b). This measure comprised the listening, television, and reading comprehension variables.

Results

Preliminary Analyses

Preliminary inspection of the data with skewness and kurtosis indices did not suggest major deviations from normality. The means and standard deviations of all measures used are presented in Table 1.

Structural Equation Modeling

We used structural equation modeling (SEM) to explore the relations between oral language and decoding skills across development. Separate models were fitted for each cohort. In the evaluation of the goodness of fit of each model to the data, we report the model chi-square statistic associated with the p value, the comparative fit index (CFI), the nonnormed fit index (NNFI), the Akaike's information criterion (AIC), and the root-mean-square error of approximation (RMSEA). A nonsignificant value of the chi-square statistic indicates a good fit; however, the test is sensitive to sample size and should be considered in relation to its degrees of freedom (i.e., dividing chi-square value by its degrees of freedom should result in a value below 2, indicating a good model fit; Maruyama, 1998). CFI and NNFI indices equal to or superior to .95 are considered to indicate a good fit (Hu & Bentler, 1999). The AIC measure indicates a better fit when it is smaller (Browne & Cudeck, 1992). Finally, the RMSEA is an absolute fit

Table 1
Descriptive Statistics of Assessment Measures

Measure	Preschool (4 years old)				Kindergarten (6 years old)				Second grade (8 years old)			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Cohort 1												
Oral language skills												
Listening comp	4.08	2.83	0	16	5.43	3.75	0	20				
Television comp	6.54	4.17	0	18	7.44	4.44	0	28				
Decoding skills												
Letter ID	17.01	10.82	0	38	36.01	4.17	21	46				
Word ID	0.82	5.07	0	40	24.68	19.16	0	81				
DIBELS	8.76	3.97	0	16	15.32	1.20	9	16				
PPVT–III	73.81	16.09	9	114	105.23	13.74	75	147				
Cohort 2												
Oral language skills												
Listening comp					9.62	5.78	0	28	10.26	5.08	1	30
Television comp					10.79	6.26	1	36	13.13	7.01	2	37
Decoding skills												
Letter ID					34.59	5.48	4	43	—	—	—	—
Word ID					18.51	20.28	0	76	65.10	10.53	36	93
DIBELS					14.71	1.88	4	16	—	—	—	—
PPVT–III					105.01	12.95	77	143	128.76	14.91	93	167
Reading comp					—	—	—	—	13.29	5.99	0	24

Note. Raw scores are included. Comp = comprehension; ID = identification; DIBELS = Dynamic Indicators of Basic Early Literacy Skills; PPVT–III = Picture Vocabulary Test (3rd ed.).

index in which the complexity of the model is considered; values less than .05 are considered a good fit (Cudeck & Browne, 1992).

Fitting the Model for the First Cohort (Ages 4–6)

For the first cohort of children at both testing times (at ages 4 and 6), indicators for the decoding skills latent variable were phonological awareness, letter identification, and word identification, whereas indicators for the oral language skills were listening comprehension, television comprehension, and vocabulary. Figure 2 depicts the fitted model for the first cohort of children with standardized parameter estimates. We hypothesized that oral language and decoding skills relate to each other within a year. Because the development of oral language skills precedes that of decoding skills, we hypothesized that in preschool and in kindergarten, oral language skills predict decoding skills. We also hypothesized longitudinal continuity, in that preschool oral language skills predict kindergarten oral language skills and preschool decoding skills predict kindergarten decoding skills. As indicated by the fit indices, $\chi^2(49, N = 113) = 56.44$, $p = .22$ ($\chi^2/df = 1.15$); CFI = .98; NNFI = .97; AIC = 114.44; RMSEA = .04 (90% confidence interval [CI]: 0.0, 0.07), the model yielded a good fit to the data. The relative magnitude and significance of the standardized coefficients show that the relation between decoding and oral language skills is strong in preschool (with oral language skills predicting 28% of the variance in de-

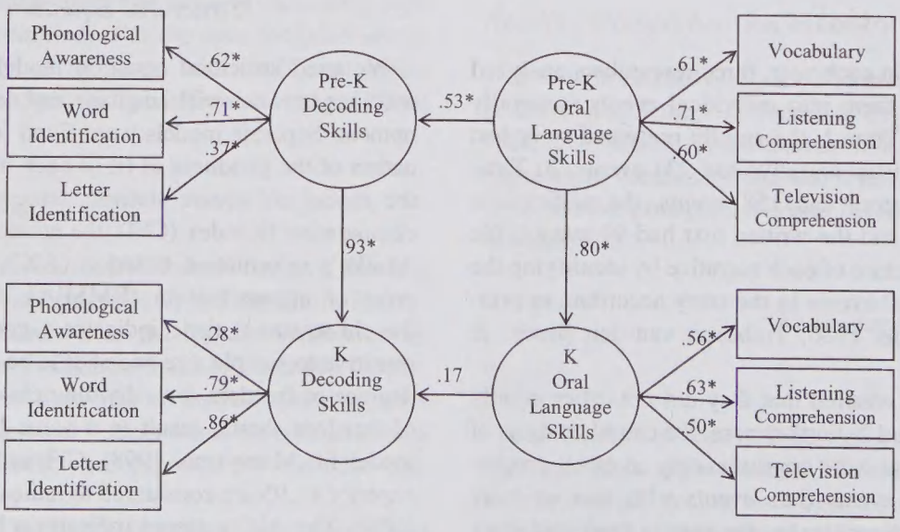


Figure 2. Model for Cohort 1 (ages 4–6) depicting relations between decoding and oral language skills.

coding skills) and very weak (not significant) in kindergarten. Furthermore, there is longitudinal continuity within both oral language and decoding skills. Specifically, for oral language skills, approximately 64% of the variance in kindergarten children's ability was explained by their ability 2 years earlier in preschool. Likewise, for decoding skills, approximately 75% of the variance in kindergarten children's ability was explained by their ability 2 years earlier in preschool.

Fitting the Model for the Second Cohort (Ages 6–8)

For the second cohort at Time 1 (age 6), indicators for the decoding skills latent variable were letter identification, word identification, and phonological awareness. Indicators for oral language skills were listening comprehension, television comprehension, and vocabulary. At Time 2 (age 8), decoding skills were estimated by a single variable, word identification; indicators for the oral language latent variable were listening comprehension, television comprehension, and vocabulary. Reading comprehension also was estimated by a single variable. Figure 3 depicts the fitted model for the second cohort of children with standardized parameter estimates. Consistent with the model fitted for the first cohort, we hypothesized that kindergarten and second grade oral language skills precede and, therefore, predict decoding skills. We also hypothesized longitudinal continuity in that kindergarten oral language skills predict second grade oral language skills and kindergarten decoding skills predict second grade decoding skills. In addition, we hypothesized that both oral language and decoding skills in second grade independently predict reading comprehension in second grade. As indicated by the fit indices, $\chi^2(36, N = 108) = 38.03$, $p = .38$ ($\chi^2/df = 1.06$); CFI = .99; NNFI = .99; AIC = 98.03; RMSEA = .02 (90% CI: 0.0, 0.07), the model yielded a good fit to the data.

The relative magnitude and significance of the standardized coefficients show that the relation between decoding and oral language skills is very weak and not significant in either kindergarten or second grade. Furthermore, there is longitudinal continuity within decoding skills; approximately 63% of the variance in a child's ability in second grade was accounted for by his or her ability in kindergarten. Longitudinal continuity within the oral language skills is rather weak; approximately only 3% of the variance in a second grade child's ability was significantly explained by his or her ability in kindergarten. Of note, reading comprehension in second grade (age 8) is influenced by both decoding (i.e., word identification) and oral language skills. Specifically, approximately 47% of the variance in reading comprehension is explained by oral language and decoding skills.

Testing Alternative Models

The hypothesized models fit the data well, as suggested by the goodness-of-fit measures. However, it is possible that our data support other models that are also theoretically meaningful. To explore this hypothesis, we tested a series of alternative models for each cohort and compared them with the models in Figures 2 and 3. When the alternative and original models were nested (i.e., they included the same number of parameters), we evaluated their overall fits based on traditional fit indices and compared them on the basis of overall fit, the discrepancy/degrees of freedom ratio (χ^2/df), the AIC, and a chi-square test. The discrepancy/degrees of freedom ratio of 2 or less indicates a close fit, whereas the AIC measure indicates a better fit when it is smaller (Browne & Cudeck, 1992). With the chi-square test, we tested the null hypothesis of no significant difference in fit by evaluating whether the chi-square difference between the two models is significant for the given degrees of freedom and a chosen significance level. If the

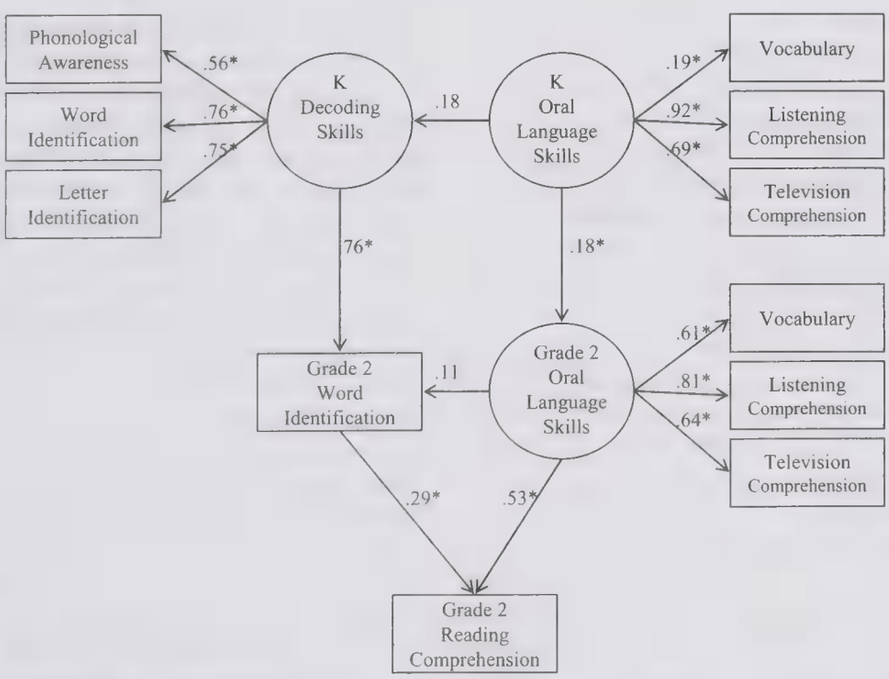


Figure 3. Model for Cohort 2 (ages 6–8) depicting relations between decoding skills, oral language skills, and reading comprehension. *Note.* Values are standardized coefficients. $^*p < .05$.

difference is significant, then the null hypothesis is rejected, and significant differences between the models are indicated. When the alternative and original models were not nested, we evaluated overall fits based on traditional fit indices and compared them on the basis of overall fit, the discrepancy/degrees of freedom ratio (χ^2/df), and the AIC.

The first alternative model that we tested for the first cohort differed from the original model in Figure 2 in that it eliminates the within-age-group relations between decoding and oral language skills in each year. Thus, this model assumes complete independence between decoding and oral language skills (Figure 4, Model A). The overall fit of this model was not acceptable, $\chi^2(51, N = 113) = 70.67$, $p = .03$ ($\chi^2/df = 1.38$); CFI = .94; NNFI = .92; AIC = 124.67; RMSEA = .06 (90% CI: 0.02, 0.09). This fit is significantly weaker than that of the original model, $\Delta\chi^2(2) = 14.24$, $p < .05$. Following the same rationale, we tested the same alternative model (i.e., without the within-age-group relations between decoding and oral language skills) for the second cohort (Figure 4, Model B). The overall fit of this model was not acceptable, $\chi^2(40, N = 108) = 66.34$, $p = .004$ ($\chi^2/df = 1.66$); CFI = .92; NNFI = .89; AIC = 128.34; RMSEA = .08 (90% CI: 0.0, 0.08). This fit is significantly weaker than that of the original model, $\Delta\chi^2(4) = 28.31$, $p < .05$. In summary, this alternative model significantly decreased in fit for both cohorts when compared with the original models depicted in Figures 2 and 3. On this ground, we accepted the original models.

The second alternative model that we tested for the first cohort differed from the original model in that it includes longitudinal relations between decoding and oral language skills. In this model, we assumed relations between decoding and oral language skills within and across years (Figure 4, Model C). The overall fit of this model was good, $\chi^2(47, N = 113) = 55.77$, $p = .18$ ($\chi^2/df = 1.18$); CFI = .97; NNFI = .96; AIC = 117.77; RMSEA = .04 (90% CI: 0.0, 0.08). In this model, the added path from prekindergarten oral language to kindergarten decoding skills was significant (.22, $p < .05$), whereas the path from prekindergarten decoding skills to kindergarten oral skills was not significant (.10, $p > .05$). The chi-square test suggested that the two models were not significantly different, $\Delta\chi^2(2) = .66$, $p > .05$; however, the original model had smaller RMSEA, AIC, and χ^2/df. Following the same rationale, we tested this alternative model—that is, with the addition of the longitudinal relations between decoding and oral language skills—for the second cohort (Figure 4, Model D). The overall fit of this model also was good, $\chi^2(34, N = 108) = 35.63$, $p = .39$($\chi^2/df = 1.05$); CFI = .99; NNFI = .99; AIC = 99.63; RMSEA = .02 (90% CI: 0.0, 0.07). Both paths added in this model, from kindergarten oral language to second grade decoding skills (.14, $p > .05$) and from kindergarten decoding skills to second grade oral skills (.08, $p > .05$), were not significant. The chi-square test suggested that the two models were not significantly different, $\Delta\chi^2(2) = 2.40$, $p > .05$; however, the original model had smaller AIC and χ^2/df. In summary, for both cohorts this alternative model and the original model had a good fit to the data. However, the original model is less complex than the alternative model. The principle of parsimony suggests adopting the model with fewer estimated parameters (Marsh & Hau, 1996), namely the original model for each cohort depicted in Figures 2 and 3.

The third alternative model that we tested for the first cohort eliminated the within-age-group relations between decoding and oral language skills and added longitudinal relations. In this model, we assumed independence between decoding and oral skills within years but not across years (Figure 4, Model E). The overall fit of this model was not acceptable, $\chi^2(49, N = 113) = 68.29$, $p = .03$ (ratio $\chi^2/df = 1.40$); CFI = .94; NNFI = .92; AIC = 126.29; RMSEA = .06 (90% CIs: 0.10, 0.09). This fit is weaker than that of the original model. Following the same rationale, we tested this alternative model for the second cohort (Figure 4, Model F). The overall fit of this model was good, $\chi^2(36, N = 108) = 41.28$, $p = .25$ (ratio $\chi^2/df = 1.15$); CFI = .97; NNFI = .98; AIC = 101.28; RMSEA = .04(90% CI: 0.0, 0.08); however, the original model had smaller RMSEA, AIC, and χ^2/df. In summary, this alternative model significantly decreased in fit for both cohorts when compared with the original model. On this ground, we accepted the original model for each cohort depicted in Figures 2 and 3.

Testing Sensitivity to the Causal Structure as a Comprehension Measure

On the basis of strong evidence in the literature, we used sensitivity to causal structure as the primary, theoretically derived way of assessing children's comprehension of stories and of tracking their development. Accordingly, in assessing children's comprehension, we used recall of highly connected events in each story and not simply overall recall. The assumption here was that recall of highly connected events is a better indicator of children's comprehension than overall recall, even though the two measures often are highly interrelated. Indeed, correlations in both cohorts ranged from .89 to .97. To test this hypothesis, for each cohort we formulated an alternative model that included overall recall as indicators for listening and for television and reading comprehension and compared it with the original model in Figures 2 and 3. The alternative and original models were evaluated on the basis of traditional fit indices and compared on the basis of overall fit, the discrepancy/degrees of freedom ratio (χ^2/df), and the AIC.

The alternative model to the original model in Figure 2 for the first cohort included overall recall of the listening and the television comprehension in addition to vocabulary as indicators for the oral skills in prekindergarten and kindergarten (instead of the highly connected recall). The overall fit of this model was good, $\chi^2(49, N = 113) = 62.12$, $p = .09$ ($\chi^2/df = 1.27$); CFI = .96; NNFI = .95; AIC = 120.12; RMSEA = .05 (90% CI: 0.0, 0.08), but weaker than that of the original model; the original model had smaller RMSEA, .04, $\chi^2/df = 1.15$, and AIC = 68.44. The alternative model to the original model in Figure 3 for the second cohort included overall recall of the listening and television and reading comprehension as indicators (instead of the highly connected recall). The overall fit of this model was good, $\chi^2(36, N = 108) = 46.45$, $p = .11$ (ratio $\chi^2/df = 1.29$); CFI = .97; NNFI = .96; AIC = 106.45; RMSEA = .05 (90% CI: 0.0, 0.09), but weaker than that of the original model; the original model had smaller RMSEA, .02, $\chi^2/df = 1.06$, and AIC = 76.03. In summary, the alternative model for each cohort slightly decreased in fit when compared with the original models in Figures 2 and 3. On this ground, we accepted the original model in Figures 2 and 3.

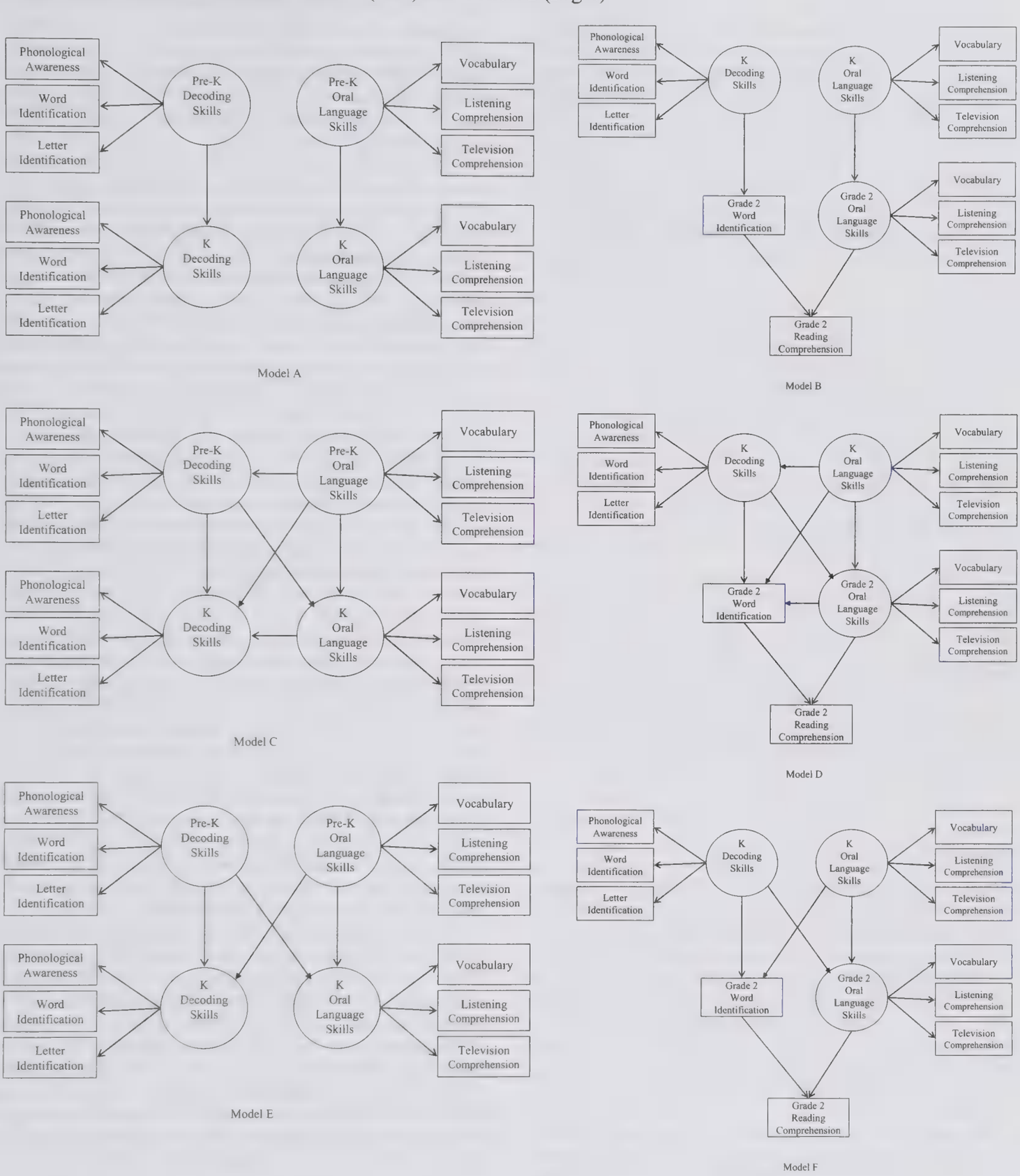


Figure 4. Alternative models tested for Cohort 1 (left) and Cohort 2 (right).

Cohort Comparison

In this study, data were collected twice from 6-year-old children. The 6-year-old children in the first cohort (at Time 2) and the 6-year-old children in the second cohort (at Time 1) were administered the same decoding and vocabulary measures but different aural and television narratives. To evaluate the degree to which the two cohorts of children differed in their performance, we directly compared their scores on the three decoding and three oral skill measures. To do so, we conducted a multivariate analysis of variance with cohort as the independent variable and phonological awareness, letter identification, word identification, listening comprehension, television comprehension, and vocabulary as the dependent variables. This analysis revealed significant cohort differences, $F(6, 214) = 10.99$, $p = .0001$, $\eta^2 = .24$. First, with respect to decoding skills, the first cohort (6 years old at Time 2) performed significantly higher than the second cohort (6 years old at Time 1) in phonological awareness, $F(1, 219) = 8.89$, $p = .003$, $\eta^2 = .04$; letter identification, $F(1, 219) = 4.74$, $p = .03$, $\eta^2 = .03$; and word identification $F(1, 219) = 5.39$, $p = .02$, $\eta^2 = .02$ (all means are provided in Table 1). Second, with respect to oral language skills, the first cohort (6 years old at Time 2) performed significantly lower than the second cohort (6 years old at Time 1) in listening comprehension, $F(1, 219) = 41.12$, $p = .0001$, $\eta^2 = .16$, and television comprehension, $F(1, 219) = 21.21$, $p = .0001$, $\eta^2 = .09$. There were no differences between cohorts in vocabulary, $F(1, 219) < 1$, $p > .05$.

These cohort differences suggest that the first cohort performed higher on decoding skills than the second cohort, whereas the second cohort performed higher on oral language skills than the first cohort. Although these cohort differences emerged, the models for the two cohorts showed remarkable consistency with respect to the relations between decoding and oral language skills. In both cohorts at age 6, the standardized coefficients linking oral language and decoding skills were not significant and were of comparable magnitude (.17 and .18 for Cohorts 1 and 2, respectively).

Discussion

The present study had two goals. The first goal was to examine the relation between the development of children's oral language skills and decoding skills from preschool to early elementary school. The second goal was to examine the stability and predictive power of children's oral language skills for later reading comprehension to determine whether oral language skills at an early age predicted those at a later age and whether oral language skills were predictive of reading comprehension at a later age in early elementary school. To achieve these goals, we examined oral language and decoding skills of two cohorts of children, 4 and 6 years old, and retested them when they were 6 and 8 years old, respectively.

The findings show, first, that within each age group, oral language skills and decoding skills formed distinct clusters and that these clusters showed longitudinal continuity. With regard to the latter finding, SEM showed that oral language skills at one age uniquely predicted oral language skills 2 years later for both cohorts of children (i.e., from age 4 to age 6, and from age 6 to age 8) and, likewise, that decoding skills at an early age uniquely predicted decoding skills 2 years later. Second, the findings showed that the two sets of skills were strongly interrelated in preschool and that with development this relation became weaker, both from preschool to kindergarten and from kindergarten to second grade. Specifically, SEM showed that in preschool, oral language skills predicted decoding skills but that this pattern was weaker in kindergarten and second grade. Third, oral language and decoding skills each independently predicted a child's reading comprehension in second grade.

These conclusions are based on the original model fitted for each cohort hypothesizing that oral language skills predict decoding skills within each year. To ascertain the validity of this model, we considered several plausible alternative models. One set of models hypothesized complete independence of decoding and oral language skills within and across years (Models A and B for the respective cohorts). Another set of models hypothesized that oral language and decoding skills predict each other within and across years (Models C and D). A third set of models hypothesized that oral language and decoding skills predict each other only across years (Models E and F). These alternative models either had poorer fit to the data or, in the case of Models C and D, had a good fit to the data but were less parsimonious than the original models.

Given that the principle of parsimony is a methodological criterion, it is worthwhile to consider the possible implications of adopting a less parsimonious, theoretically viable model that has good fit to the data. In this model (Models C and D for the respective cohorts), oral language skills not only predict decoding skills *within* a year but oral language skills at a younger age also predict decoding skills at a later age, that is, *across* years. Although these two sets of skills follow their own developmental trajectories across time as indicated by the dissociation between oral language and decoding skills that has been reported for older children (Oakhill, Cain, & Bryant, 2003; Paris & Paris, 2001, 2003; van den Broek, et al., 1996; Vellutino et al., 2007; Whitehurst & Lonigan, 1998), the current findings extend that observation by showing that distinct clusters of oral language skills and decoding skills are present at a much earlier age than previously shown and that they have a reciprocal relation: The development of one influences directly the development of the other and vice versa.

With respect to reading comprehension, the current results confirm a time-honored view showing that the two clusters of skills contribute to a child's reading comprehension activities in early elementary school, with each cluster making a sizable, unique contribution. Thus, the particular requirements of comprehending a text involve both the ability to "break the code" by translating written symbols into meaningful words (Adams, 1990; Ehri, 1999, 2005; Perfetti, 1985; Stanovich, 1986) and the ability to extract meaning about events and facts and identify semantic relations between those events and facts (cf. Ehri, 1998; Graesser et al., 1994; Kintsch, 1988; Ruddell & Ruddell, 1994; van den Broek & Kremer, 1999; Vellutino et al., 2007; Whitehurst & Lonigan, 1998). Our results show that both types of skills begin developing during the preschool years and that these early skills are predictive of reading comprehension in second grade (see also Cain & Oakhill, 2007; Paris & Paris, 2003).

In light of the finding that both oral language and decoding skills predict reading comprehension, one would expect the impact of decoding skills to gradually diminish as elementary school children become more proficient, with increasing amounts of vari-

ance in successful reading comprehension being contributed by oral language skills. Indeed, that is what we observed in our study. Oral language skills accounted for more variance in reading comprehension than did decoding skills. Likewise, when the contributions of skills such as these to reading comprehension were assessed in second/third grade (i.e., an age similar to that of the oldest group in our study) and in sixth/seventh grade, the contribution of oral language skills was found to increase across age (Vellutino et al., 2007).

In this study, comprehension was operationalized as children's sensitivity to the causal structure of the narrative. Such operationalization has considerable psychological validity (Graesser & Clark, 1985; Trabasso et al., 1989; van den Broek et al., 1996; van den Broek, 1989a) and reflects the fact that creating coherence is at the center of successful oral and reading comprehension (Graesser et al, 1994; Kintsch, 1988; Oakhill & Cain, 2007; Rapp, van den Broek, McMaster, Kendeou, & Espin, 2007). This measure accounts for qualitative as well as quantitative developmental aspects of children's comprehension skills, and our findings suggest that it estimates children's comprehension skills better than measures that are simply focused on overall recall. We suggest that sensitivity to the causal structure provides an alternative to the passage comprehension subsections of standardized tests frequently used in prior studies. Such tests have been designed for students who have mastered decoding skills and are widely criticized as invalid measures of comprehension (e.g., Cutting & Scarborough, 2006; Fletcher, 2006; Hannon & Daneman, 2001; Keenan & Betjemann, 2006).

These findings have important theoretical and practical implications. With respect to theoretical implications, they show that successful reading comprehension depends on decoding skills, such as phonological awareness and letter and word identification, but also on oral language skills such as vocabulary and discourse comprehension. Both sets of skills make unique contributions to reading comprehension when children become conventional readers. These findings extend to the earlier preschool-level findings on school-age children (Catts, Hogan, & Fey, 2003; Nation & Snowling, 1997, 1998; Oakhill et al., 2003; Paris & Paris, 2001, 2003; Savage, 2006; Tunmer & Hoover, 1992), and moreover, they do so by following individual children in a cross-sequential longitudinal design rather than by means of cross-sectional designs. The two sets of skills begin to develop early in children's lives, at first intertwined but gradually developing relatively independently, with each having considerable stability over development. In the early elementary grades, these sets of skills combine to support children's reading comprehension. Thus, instead of a single causal sequence of skill development, the reality is one of multiple causality, with the two different sets of skills both being necessary and neither alone sufficient for later reading comprehension success (van den Broek et al., 2005).

From a practical point of view, these findings have implications for comprehension assessment, intervention, and direct instruction. First, they show that early assessment of comprehension skills is not only possible but also useful because the results predict future comprehension performance in reading contexts, namely when children are engaged in reading activities (Kendeou et al., 2005, 2007; Paris & Stahl, 2005; van den Broek et al., 2005). The overlap between comprehension performances in different media indicates that comprehension assessment in nonreading contexts (e.g., listening and television comprehension) may be used for early identification of students who are likely to experience later difficulties in reading comprehension.

It is important to note that the assessment of comprehension skills focusing on children's sensitivity to the causal structure not only is supported by our data but also accounts for qualitative as well as quantitative developmental aspects of children's comprehension skills. In this assessment, we focused not only on the *number* of the connections in individuals' representations as a function of the story structure but also of the *types* of connections included. Whereas one sign of improved comprehension is that a child identifies more relations in a story, another—and perhaps more telling—sign is that a child has advanced to include complex types of relations (such as causal connections).

Second, with regard to comprehension interventions, the findings suggest that improvement and development of comprehension skills in preschool children may lead to improved comprehension once those children reach reading age. As with assessment, such interventions could rely on nonreading (e.g., televised or listening) contexts. For example, activities around television viewing or listening may provide the opportunity for developing and fostering comprehension strategies even before the beginning of formal instruction or for older, struggling readers who experience decoding difficulties.

Third, with regard to instruction, the dissociation of oral language and decoding skills has direct educational implications as it provides a conceptual framework for designing appropriate teaching practices that target both decoding and comprehension skills (e.g., Aaron, 1991; Kendeou et al., 2005, 2007; McNamara, 2007; Oakhill et al., 2003; Savage, 2001, 2006). For instance, the separation of these skills enables teachers to understand what they need to teach about decoding and comprehension within a broad curriculum. For example, in the United Kingdom, the simple view of reading has been adopted as the theoretical basis of the revised national curricular advice to all schools in England (Kendeou et al., in press; Rose, 2006; Stuart, Stainthorp, & Snowling, 2008; U.K. Department for Education and Skills, 2006).

Successful reading comprehension is the result of a confluence of elemental skills, each of which has its own developmental trajectory. The trajectories may intertwine and influence each other at early stages (e.g., in the 4-year-old children in our study), but they also remain independent to a considerable degree. The risk of a child developing reading comprehension difficulties is smallest when he or she progresses appropriately along each trajectory. The more educational practice can help each child move forward along each dimension, the more it ensures the child against failure. This means that decoding skill development should be part of the curriculum—as it traditionally has—but so should oral language skill development including narrative comprehension.

References

Aaron, P. G. (1991). Can reading disabilities be diagnosed without using intelligence tests? *Journal of Learning Disabilities, 24,* 178–186.

Ackerman, B. P. (1988). Thematic influences on children's judgments about story adequacy. *Child Development, 59,* 918–938.

Adams, M. J. (1990). *Beginning to read.* Cambridge, MA: MIT Press.

Applebee, A. N. (1978). *The child's concept of a story: Ages two to seventeen.* Chicago: University of Chicago Press.

Bauer, P. J. (1996). What do infants recall of their lives? Memory for specific events by 1- to 2-year-olds. *American Psychologist, 51,* 29–41.

Bauer, P. J. (1997). Development of memory in early childhood. In N. Cowan (Ed.), *The Development of Memory in Childhood* (pp. 83–111). Sussex, UK: Psychology Press.

Bishop, D. V. M., & Adams, C. (1990). A prospective study of the relationship between specific language impairment, phonological disorders and reading retardation. *Journal of Child Psychology and Psychiatry, 31,* 1027–1050.

Bourg, T., Bauer, P., & van den Broek, P. (1997). Building the bridges: The development of event comprehension and representation. In P. van den Broek, P. Bauer, & T. Bourg (Eds.), *Developmental spans in event comprehension and representation: Bridging fictional and actual events* (pp. 385–407). Hillsdale, NJ: Erlbaum.

Brown, A. L., & Smiley, S. S. (1978). The development of strategies for studying texts. *Child Development, 49,* 1076–1088.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21,* 230–258.

Bryant, P., MacLean, M., & Bradley, L. (1990). Rhyme, language, and children's reading. *Applied Psycholinguistics, 11,* 237–252.

Cain, K., & Oakhill, J. (2007). Reading comprehension difficulties: Correlates, causes, and consequences. In Cain, K. & Oakhill, J. (Eds). *Children's comprehension problems in oral and written language: A cognitive perspective* (pp. 41–75). New York: Guilford Press.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96,* 31–42.

Casteel, M. (1993). Effects of inferences necessity and reading goal on children's inferential integration. *Journal of Educational Psychology, 88,* 484–507.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, B. J. (1999). Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading, 3,* 331–361.

Catts, H. W., Hogan, T. P., & Fey, M. E. (2003). Subgrouping poor readers on the basis of individual differences in reading-related abilities. *Journal of Learning Disabilities, 36,* 151–164.

Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika, 57,* 357–369.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10,* 277–299.

Duke, N. K. (2005). Comprehension of what for what: Comprehension as a non-unitary construct. In S. Paris & S. Stahl (Eds.), *Current issues in reading comprehension and assessment* (pp. 93–104). Mahwah, NJ: Erlbaum.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test (3rd ed.) (PPVT–III).* Circle Pines, MN: American Guidance Services.

Ehri, L. C. (1998). Word reading by sight and by analogy in beginning readers. In C. Hulme & R. M. Yoshi (Eds.), *Reading and spelling: Development and disorders* (pp. 87–111). London: Erlbaum.

Ehri, L. C. (1999). Phases of development in learning to read words. In J. Oakhill & R. Beard (Eds.), *Reading development and the teaching of reading: A psychological perspective* (pp. 79–108). Oxford, UK: Blackwell.

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading, 9,* 167–188.

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research, 71,* 393–447.

Fletcher, J. M. (2006). Measuring reading comprehension. *Scientific Studies of Reading, 10,* 323–330.

Francis, D. J., Fletcher, J. M., Catts, H., & Tomblin, B. (2005). Dimensions affecting the assessment of reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 369–394). Mahwah, NJ: Erlbaum.

Francis, D. J., Snow, C. E., August, D., Carlson, C. D., Miller, J., & Iglesias, A. (2006). Measures of reading comprehension: A latent variable analysis of the diagnostic assessment of reading comprehension. *Scientific Studies of Reading, 10,* 301–322.

Gernsbacher, M. A. (1990). *Language comprehension as a structure building.* Hillsdale, NJ: Erlbaum.

Goldman, S. R., & Varnhagen, C. K. (1986). Memory for embedded and sequential story structures. *Journal of Memory and Language, 25,* 401–418.

Good, R. H., & Kaminski, R. A. (2002a). *DIBELS oral reading fluency passages for first through third grades.* (Technical Report No. 10). Eugene, OR: University of Oregon.

Good, R. H., & Kaminski, R. A. (2002b). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement. Retrieved April 12, 2006, from http://dibels.uoregon.edu/

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7,* 6–10.

Graesser, A., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative comprehension. *Psychological Review, 101,* 371–395.

Graesser, A. C., & Clark, L. F. (1985). *The Structures and Procedures of Implicit Knowledge.* Norwood, NJ: Ablex.

Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology, 93,* 103–128.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2,* 127–160.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the gray oral reading test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading, 10,* 363–380.

Kendeou, P., Lynch, J. S., van den Broek, P., Espin, C. A., White, M. J., & Kremer, K. E. (2005). Developing successful readers: building early comprehension skills through television viewing and listening. *Early Childhood Education Journal, 33,* 91–98.

Kendeou, P., & Papadopoulos, T. C. (2009). *Reading comprehension tests: What skills do they assess?* Manuscript in preparation.

Kendeou, P., Savage, R. S., & van den Broek, P. (in press). Revisiting the simple view of reading. *British Journal of Educational Psychology.*

Kendeou, P., van den Broek, P., White, M., & Lynch, J. (2007). Preschool and early elementary comprehension: Skill development and strategy interventions. In D. S. McNamara (Ed.) *Reading comprehension strategies: Theories, interventions, and technologies,* (pp. 27–45). Mahwah, NJ: Erlbaum.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95,* 163–183.

Kintsch, W., & van Dijk, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review, 85,* 363–394.

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study. *Developmental Psychology, 36,* 596–613.

Lorch, E., Bellack, D., & Augsbach, L. (1987). Young children's memory for televised stories: Effects of importance. *Child Development, 58,* 453–463.

Lorch, E. P., & Sanchez, R. P. (1997). Children's memory for televised

events. In P. van den Broek, P. J. Bauer, & T. Bourg (Eds.), *Developmental spans in event comprehension and representation* (pp. 271–291). Mahwah, NJ: Erlbaum.

Lynch, J. S., van den Broek, P., Kremer, K. E., Kendeou, P., White, M., & Lorch, E. P. (2008). The development of narrative comprehension in its relation to other early reading skills. *Reading Psychology, 29,* 327–365.

Mackie, J. L. (1980). *The cement of the universe: A study of causation.* Oxford, UK: Clarendon.

Magliano, J. P., Millis, K. K., Ozuru, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 107–136). Mahwah, NJ: Erlbaum.

Mandler, J. M. (1984). *Stories, scripts, and scenes: Aspects of schema theory.* Hillsdale, NJ: Erlbaum.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology, 9,* 111–151.

Mares, M. L. (2006). Repetition increases children's comprehension of television content—Up to a point. *Communication Monographs, 73,* 216–241.

Marsh, H. W., & Hau, K. T. (1996). Assessing goodness of fit: Is parsimony always desirable? *The Journal of Experimental Education, 64,* 364–390.

Maruyama, G. M. (1998). *Basics of structural equation modeling.* Thousand Oaks, CA: Sage Publications.

McNamara, D. S. (Ed.). (2007). *Reading comprehension strategies: Theory, interventions, and technologies.* Mahwah, NJ: Erlbaum.

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67,* 359–370.

Nation, K., & Snowling, M. (1998). Semantic processing and the development of word-recognition skills: Evidence from children with reading comprehension difficulties. *Journal of Memory and Language, 39,* 85–101.

O'Brien, E. J., & Myers, J. L. (1987). The role of causal connections in the retrieval of text. *Memory and Cognition, 15,* 419–427.

Oakhill, J., & Cain, K. (2007). Issues of causality in children's reading comprehension. In McNamara, D. (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 47–72). New York: Erlbaum.

Oakhill, J. V., Cain, K., & Bryant, P. E. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes, 18,* 443–468.

Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: The passage or the question? *Behavior Research Methods, 40,* 1001–1015.

Papadopoulos, T. C., Das, J. P., Parrila, R. K., & Kirby, J. R. (2003). Children at-risk for developing reading difficulties: A remediation study. *School Psychology International, 24,* 340–366.

Paris, A. H., & Paris, S. G. (2003). Assessing narrative comprehension in young children, *Reading Research Quarterly, 38,* 36–76.

Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist, 36,* 89–101.

Paris, S. G., & Stahl, S. A. (Eds.). (2005). *Children's reading comprehension and assessment.* Mahwah, NJ: Erlbaum.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices-past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13–69). Mahwah, NJ: Erlbaum.

Perfetti, C. A. (1985). *Reading ability.* New York: Oxford University Press.

Pressley, M., Wharton-McDonald, R., Allington, R., Block, C. C., Morrow, L., Tracey, D., et al. (2001). A study of effective first-grade literacy instruction. *Scientific Studies of Reading, 5,* 35–58.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension.* Santa Monica, CA: RAND Corp.

Rapp, D. N., van den Broek, P., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading, 11,* 289–312.

Rose, J. (2006). *Independent review of the teaching of early reading: Final report.* London: Department for Education and Skills/TSO.

Ruddell, R. B., & Ruddell, M. R. (1994). Language acquisition and literacy process. In R. B. Ruddell, M. R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading,* (4th ed., pp. 83–103). Newark, DE: International Reading Association.

Savage, R. S. (2001). The "simple view" of reading: Some evidence and possible implications. *Educational Psychology in Practice, 17,* 17–33.

Savage, R. S. (2006). Reading comprehension is not always the product of nonsense word decoding and linguistic comprehension: Evidence from teenagers who are extremely poor readers. *Scientific Studies of Reading, 10,* 143–164.

Snowling, M. J., & Hulme, C. (2005). *The science of reading: A handbook.* Malden, MA: Blackwell.

Speece, D. L., Roth, F. P., Cooper, D. H., & de la Paz, S. (1999). The relevance of oral language skills to early literacy: A multivariate analysis. *Applied Psycholinguistics, 20,* 167–190.

Stanovich, K. E. (1986). Cognitive processes and the reading problems of learning-disabled children: Evaluating the assumption of specificity. In B. Y. L. Wong & J. K. Torgeson (Eds.), *Psychological and educational perspectives on learning disabilities* (pp. 87–131). Orlando, FL: Academic Press.

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In R. 0. Freedle (Ed.), *New directions in discourse processing: Vol. 2. Advances in discourse processes.* (pp. 53–120). Hillsdale, NJ: Erlbaum.

Storch, S., & Whitehurst, G. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology, 38,* 934–947.

Stuart, M., Stainthorp, R., & Snowling, M. (2008). Literacy as a complex activity: Deconstructing the simple view of reading. *Literacy, 42,* 59–66.

Trabasso, T., & Nickels, M. (1992). The development of goal plans of action in the narration of a picture story. *Discourse Processes, 15,* 249–275.

Trabasso, T., Secco, T., & van den Broek, P. (1984). Causal cohesion and story coherence. In H. Mandl, N. L. Stein, & T. Trabasso (Eds.), *Learning and comprehension of text* (pp. 83–111). Hillsdale, NJ: Erlbaum.

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language, 24,* 612–630.

Trabasso, T., van den Broek, P., & Liu, L. (1988). A model for generating questions that assess and promote comprehension. *Questioning Exchange, 2,* 25–38.

Trabasso, T., van den Broek, P. W., & Suh, S. Y. (1989). Logical necessity and transitivity of causal relations in stories. *Discourse Processes, 12,* 1–25.

Tunmer, W. E., & Hoover, W. A. (1992). Cognitive and linguistic factors in learning to read. In P. Gough, L. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 175–214). Hillsdale, NJ: Erlbaum.

U.K. Department for Education and Skills. (2006). *Five-year strategy for children and learners: Maintaining the excellent progress. Strategy paper.* London: TSO.

van den Broek, P. (1990). The causal inference maker: Towards a process model of inference generation in text comprehension. In D. A. Balota, G. B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 423–446). Hillsdale, NJ: Erlbaum.

van den Broek, P. (1994). Comprehension and memory of narrative texts.

In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 539–588). London: Academic Press.

van den Broek, P. (1997). Discovering the cements of the universe: The development of event comprehension from childhood to adulthood. In P. van den Broek, P. Bauer, & T. Bourg (Eds.), *Developmental spans in event comprehension: Bridging fictional and actual events* (pp. 321–342). Mahwah, NJ: Erlbaum.

van den Broek, P., Kendeou, P., Kremer, K., Lynch, J. S., Butler, J., White, M. J., & Lorch, E. P. (2005). Assessment of comprehension abilities in young children. In S. Stahl & S. Paris (Eds.), *Children's reading comprehension and assessment,* (pp. 107–130). Mahwah, NJ: Erlbaum.

van den Broek, P., & Kremer, K. E. (1999). The mind in action: What it means to comprehend during reading. In B. Taylor, M. Graves, & P. van den Broek (Eds.), *Reading for meaning* (pp. 1–31). New York: Teacher's College Press.

van den Broek, P., Lorch, E. P., & Thurlow, R. (1996). Children's and adults' memory for television stories: The role of causal factors, story-grammar categories, and hierarchical level. *Child Development, 67,* 3010–3028.

van den Broek, P., & Lorch, R. F. (1993). Network representations of causal relations in memory for narrative texts: Evidence from primed recognition. *Discourse Processes, 16*(1–2), 75–98.

van den Broek, P. W. (1986). Judging the importance of events in stories: The influence of structural variation and grade level. *Dissertation Abstracts International, 46*(9-B), 3243.

van den Broek, P. W. (1988). The effects of causal relations and hierarchical position on the importance of story statements. *Journal of Memory and Language, 27,* 1–22.

van den Broek, P. W. (1989a). Causal reasoning and inference making in judging the importance of story statements. *Child Development, 60,* 286–297.

van den Broek, P. W. (1989b). The effects of causal structure on the comprehension of narratives: Implications for education. *Reading Psychology: An International Quarterly, 10,* 19–44.

VanderVeen, A., Huff, K., Gierl, M., McNamara, D. S., Louwerse, M., & Graesser, A. C. (2007). Developing and validating instructionally relevant reading competency profiles measured by the critical reading sections of the SAT. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 137–172). Mahwah, NJ: Erlbaum.

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skill model of reading development. *Scientific Studies of Reading, 11,* 3–32.

Wenner, J., & Bauer, P. J. (2001). Bringing order to the arbitrary: One- to two-year-olds' recall of event sequences. *Infant Behavior and Development, 22,* 585–590.

Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development, 69,* 848–872.

Williams, J. P. (1993). Comprehension of students with and without learning disabilities: Identification of narrative themes and idiosyncratic text representations. *Journal of Educational Psychology, 85,* 631–642.

Woodcock, R. W. (1987). *Woodcock Reading Mastery Test (WRMT).* Circle Pines, MN: American Guidance Services.

Received April 8, 2008
Revision received March 11, 2009
Accepted March 27, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 779–789

0022-0663/09/$12.00 DOI: 10.1037/a0016601

Improving Classroom Learning by Collaboratively Observing Human Tutoring Videos While Problem Solving

Scotty D. Craig, Michelene T. H. Chi, and Kurt VanLehn
University of Pittsburgh

Collaboratively observing tutoring is a promising method for observational learning (also referred to as vicarious learning). This method was tested in the Pittsburgh Science of Learning Center's Physics LearnLab, where students were introduced to physics topics by observing videos while problem solving in Andes, a physics tutoring system. Students were randomly assigned to three groups: (a) pairs collaboratively observing videos of an expert human tutoring session, (b) pairs observing videos of expert problem solving, or (c) individuals observing expert problem solving. Immediate learning measures did not display group differences; however, long-term retention and transfer measures showed consistent differences favoring collaboratively observing tutoring.

Keywords: collaboration, observational learning, vicarious learning, physics, problem solving

Observing tutoring has recently emerged as a promising new focus in the observational learning literature (Chi, Roy, & Hausmann, 2008; Craig, Driscoll, & Gholson, 2004). By *observing tutoring*, we refer to the process in which a learner sees and hears the dialogue between a tutor and a tutee without being able to participate in it. In observational learning (also labeled *vicarious* or *social learning*), information is gained by watching the learning process of another (Bandura, 1986; Gholson & Craig, 2006). Thus, observing tutoring can be considered a subcategory of observational learning.

In the present study and previous research by Chi et al. (2008), pairs of students solved problems collaboratively as they observed tutoring. This combination of collaborative problem solving and observing tutoring is called *collaboratively observing tutoring*. If collaboratively observing tutoring proves to be an effective method of learning, then it could provide a cost-effective alternative to human tutoring and intelligent tutoring systems (Alessi & Trollip, 1991; Anderson, Corbett, Koedinger, & Pelletier, 1995; Azevedo, & Bernard, 1995; Derry & Potts, 1998; VanLehn et al., 2005).

However, it is important to understand whether the benefits are due to the domain content of the videos (essentially, a worked example) or the tutorial content, which has affective overtones and is conversationally based. The current study compares collaboratively observing videos of one-on-one, expert human tutoring with observing videos of an expert demonstrating how to solve the same problems. That is, the videos show either a tutoring session or worked examples.

Observational Learning

Learning by observing has been investigated in several areas of research. For instance, in social psychology, studies have shown that people who watch someone acting aggressively tend to start acting more aggressively themselves (Bandura, 1969, 1986). Neuroscientists and developmental psychologists study imitative learning in humans and other species (e.g., Meltzoff, 2005). Learning by observing occurs in work settings (Latham & Saari, 1979) and is the first stage of Collins, Brown, and Newman's (1989) model–scaffold–fade account of cognitive apprenticeship. In much of this work, students learned by observing live humans, observing videos of humans, studying cartoons, and/or listening to audiotapes (Bandura, 1986; Gholson & Craig, 2006; Rogoff, Paradise, Mejía Arauz, Correa-Chávez, & Angelillo, 2003; Rosenthal & Zimmerman, 1978).

However, when the competence to be acquired is problem solving and the observed material is a problem plus all the steps required for its solution, then the material is called a *worked example*. Typically, the problem and its steps are presented on paper, in a video, or via click-through text. Considerable research has investigated how students learn from worked examples (Atkinson, Derry, Renkl, & Wortham, 2000). Although it might be interesting to compare modalities (video vs. paper, etc.), there is no doubt that students can learn from observation in all of them. Our

Scotty D. Craig, Michelene T. H. Chi, and Kurt VanLehn, Learning Research and Development Center, University of Pittsburgh.

Michelene T. H. Chi is now in the Division of Psychology in Education, College of Education, Arizona State University, and Kurt VanLehn is now in the Department of Computer Science and Engineering, School of Computing and Informatics, Arizona State University.

This project was supported by National Science Foundation (NSF) Award SBE-0354420 to the Pittsburgh Science of Learning Center. This research was also partially supported by Institute of Education Sciences (IES) Grant R305H0R0169. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Pittsburgh Science of Learning Center (PSLC), NSF, or IES (U.S. Department of Education).

We would like to thank the Pittsburgh Science of Learning Center (PSLC) Physics LearnLab (www.LearnLab.org) for their help and support on this project. Specifically, we thank Sayaka Takeda for her help with data coding and Karl Fike for help with editing and formatting.

Correspondence concerning this article should be addressed to Scotty D. Craig, who is now in the Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152. E-mail: scraig@memphis.edu

concern is with the content presented to students, so we used video as the only presentation modality.

Whereas a common way of modeling a new skill is for the instructor to demonstrate it, another method is for the instructor to tutor a student with other students watching. For instance, a master tailor may show a senior apprentice how to sew an intricate joint while the other apprentices watch (Lave & Wenger, 1991), or a student may solve a math problem on the blackboard at the front of the class with the teacher's help. The demonstration is a monologue, whereas the tutoring session is a dialogue between the tutor and tutee.

Both types of presentation of material seem to have benefits. On the one hand, the demonstration (worked example) is probably coherent and certainly correct, whereas the tutorial dialogue may be incoherent and/or incorrect at times. From a purely cognitive point of view, the information as presented in the demonstration should be easier to learn than the information as presented in the tutoring session. On the other hand, the tutoring session might be more interesting. For instance, text has been found to have higher ratings of situational interest if it contains humans with whom the reader can identify (Hidi & Harackiewicz, 2000). The tutoring session may display self-regulatory behavior that the observers would be wise to emulate. The observers may adopt more realistic expectations for their own performance when they see another student struggle. Further, the dialogue of the tutoring session provides discourse scaffolds, such as questions and explanations, shown to be effective for learning in both classroom settings (Silliman & Wilkinson, 1994) and in vicarious learning (Craig, Brittingham, Williams, Cheney, & Gholson, 2009). In brief, the worked examples may be cognitively easier to learn from, but the tutoring sessions may provide more guidance and motivational benefits.

Although the choice between modeling with a worked example versus a tutoring session is clearly an important one that instructional designers and teachers must confront frequently, little research has been done on this issue. Early work on the dialogue versus monologue manipulation suggests that the choice can influence the learner's performance (Cox, McKendree, Tobin, Lee, & Mayes, 1999; Craig, Gholson, Ventura, Graesser, & the Tutoring Research Group, 2000; Driscoll et al., 2003; Fox Tree, 1999; Shebilske, Jordan, Goettl, & Paulus, 1998), which, in turn, suggests that the choice may influence learning. For example, Fox Tree (1999) found that performance was better while overhearing dialogues than while overhearing monologues. Fox Tree prepared materials for the experiment by dividing college students into pairs, called directors and matchers (Schober & Clark, 1989). The goal was for the director to describe an ordered set of abstract shapes (tangrams) to the matcher, so that the matcher could place the shapes in the same order as the director's pictures. Directors either gave instructions for the matcher to follow (monologue condition) or conversed freely with the matcher (dialogue condition). The sessions were recorded, and only those sessions in which matchers correctly ordered the tangrams were used as materials in the experiment. The results were that the participants who overheard dialogues outperformed those who overheard monologues on assembly tasks. The dependent measures reflected performance only (e.g., number of tangrams placed correctly) and not learning gains.

Observing tutoring has been compared with tutoring itself. Although these studies do not directly address our research question, they do indicate some factors that constrain the design of our experiment. In several experiments, Craig and colleagues (Craig et al., 2004; Craig, Sullins, Witherspoon, & Gholson, 2006) contrasted pretest to posttest gains of learners on 12 computer literacy topics. The learners either interacted directly with an intelligent tutoring system, AutoTutor (Graesser et al., 2004; Graesser, Jeon, & Dufty, 2008; Graesser, Person, Harter, & the Tutoring Research Group, 2001) or observed recordings of those tutoring sessions. Whereas learners in both conditions showed significant learning gains from pretest to posttest, participants in the computer tutoring condition significantly outperformed those in the observing tutoring condition in two experiments, with effect sizes of $d = .50$ in Study 1 and $d = .84$ in Study 2 (Craig et al., 2004). In two other experiments, there were nonsignificant trends in the same direction (Craig et al., 2006). This suggests that although computer tutoring can be more effective than observation, more testing is required to fully understand the factors impacting this finding.

Studies of learning from printed examples have shown that some students self-explain the examples and learn a great deal, whereas others read the examples in a passive way and learn considerably less (Chi, Bassok, Lewis, Reiman, & Glaser, 1989). This suggests that some observers in the Craig et al. studies may have watched the videos in a passive way that would reduce their learning.

In order to increase the number of observers using more active learning strategies, the Craig et al. (2004, Experiment 2) study was designed so that students would observe the videos in pairs. In this study, 110 participants were divided into three groups. The first two groups implemented the same computer tutoring condition ($n = 28$) and observing tutoring condition ($n = 28$) as in the other experiments. The third group consisted of pairs ($n = 27$ pairs) who observed the videos together. That is, 2 participants sat together in front of a computer monitor and watched a video of a tutoring session. They were encouraged to pause the video and talk to each other about information they did not understand in the video. Their conversation was audio recorded. All training conditions averaged about 35 min. The new condition produced gains that were intermediate between those of the tutoring condition and those of the individual observers of tutoring condition. However, the gains of the pairs in the observing tutoring condition were not significantly different from those of the two old conditions (individuals observing tutoring and tutoring).

Chi et al. (2008) compared human tutoring with observation of human tutoring along with several other control conditions. Their experiment was different from the Craig et al. (2004) experiments in several ways. First, Chi et al. used an expert tutor working face-to-face with a tutee. Second, in order to encourage equal amounts of activity, students in all conditions solved problems. The tutees solved problems with an expert tutor as a source of help. Other students solved problems with either videos of tutoring sessions or a textbook as their source of help. Third, the experiment contained five conditions: tutoring plus a 2×2 manipulation of collaboration (individual vs. pair) and source of help (textbook vs. video of tutoring sessions). Chi et al. coined the term *collaboratively observing tutoring* to refer to the condition in which pairs of students solved problems while observing a tutoring session.

The term reflects the combination of collaborative peer problem solving with observation of tutoring.

First, 10 expert tutoring sessions were conducted and video recorded. Then the other four conditions were run. Participants solved problems on paper, either individually or with a partner. While they solved problems, participants had access to either the textbook that they had studied during pretraining or to a video of the tutor and tutee solving the same problems as one they were trying to solve. Although the pairs studied together, they were assessed individually at pretest and posttest.

Chi et al. (2008) hypothesized that the combination of problem solving and collaborative problem solving would drastically reduce the frequency of passive observation of the videos and thus make collaboratively observing tutoring just as effective as face-to-face human tutoring. The predicted null effect was found. Although the number of students per condition was small ($n = 10$ for tutoring; $n = 20$ for collaboratively observing tutoring), there were statistically reliable advantages of the two conditions over the other three conditions that had similar cell sizes ($n = 20$ for pairs collaborating with a textbook; $n = 10$ for individuals observing tutoring; $n = 10$ for individuals with a textbook). These results suggest accepting the null result at face value. That is, one of the best forms of instruction known, face-to-face expert human tutoring, is no better than pairs of students solving problems while observing a video of the same problems being solved by a tutee and tutor. This intriguing result calls for further investigation and inspired the study reported here.

Taking the Craig et al. (2004) and Chi et al. (2008) studies together, one can infer a constraint on subsequent experimentation. Simply having pairs watch a video apparently does not reduce passivity nearly as much as having pairs solve a problem while they watch the tutor and the tutee solve the problem. This was confirmed in a follow-up analysis of audio recordings of the Craig et al. Experiment 2, in which pairs had an average of three conversational turns per session, with an average session lasting 35 min, whereas the collaborative observers in the Chi et al. study produced, on average, 121 conversational turns per 35-min interval. The increased level of collaboration was most likely due to the task demands of the study. Although the Craig et al. study did not require learners to perform any task other than watching the video, learners in the Chi et al. study performed a problem-solving task while observing tutoring.

Chi et al. (2008) also found that the tutoring videos were differentially effective. The observers learned more from some videos than from others. Chi et al. divided tutees into high and low pretest score groups. The pretest occurred after the pretraining and thus was partially a measure of the students' ability to learn physics. The 10 collaborative observers who viewed high-ability tutees videos learned significantly more than the 10 collaborative observers who viewed the low-ability tutee videos. This result suggests a recommendation for future experiments, specifically, the use of videos of high-ability tutees, as they somehow enhance the learning of the observers.

These studies (Chi et al., 2008; Craig et al., 2004) provide important guidelines about how to maximize *active observing* (Chi et al., 2008; Gholson & Craig, 2006) during tutoring. *Active observing* is described as observing that facilitates engagement with the materials so as to encourage deeper processing. First, observers should solve problems as they observe the video. Second, they should do so in pairs rather than working alone. And third, videos of high-ability tutees (i.e., students who have some knowledge of the material) should be used as the materials.

However, it should also be pointed out that the third most effective condition in the Chi et al. study (2008) was that of pairs collaboratively solving problems, with only a textbook as a source of help. Indeed on some measures, this group's gains were statistically equivalent to those of the top two groups (the tutees and the collaborative observers of tutoring). These gains occurred despite the fact that the textbook did not have worked examples based on the problems that the students were solving. Thus, if a pair got stuck, the pair might not have been able to find enough information in the textbook to resolve their impasse. On the other hand, if a pair in the collaboratively observing tutoring condition got stuck, they could always search the video to find out what the tutor sanctioned and resolve their impasse. Although the textbook and the tutoring videos were content equivalent in an abstract sense, the textbook lacked critical details that students might find useful when trying to solve problems and learn.

This suggested repeating the Chi et al. (2008) comparisons while controlling for the details of the content. Our experiment did so by using two kinds of videos. Both showed the same problems being solved with the same steps. However, some videos showed tutees working with a tutor to solve the problems, and the other videos showed an expert solving the problems and explaining the steps as he went. This ensured that the domain content was nearly identical. The difference in content was affective, metacognitive, and interactional. The expert-produced worked example was affect neutral, included no discussion or demonstrations of learning strategies, and, of course, contained no interaction with another person. The tutoring sessions included variations in affect, typically from the tutee; some demonstration of good and poor learning strategies (e.g., guessing, asking in-depth questions of the tutor) by the tutee; and considerable interaction between the tutor and tutee. In fact, there are probably many other differences between the two types of videos than the ones listed here. However, considerable future research is needed to identify these and to determine whether they are responsible for the learning differences we observed.

If the content in tutoring videos does increase students' active observing (e.g., interest, motivation, etc.), then pairs observing tutoring will learn more than pairs observing worked examples. By contrast, if detailed domain content is the key determinant of learning, then these two conditions should be equivalent in their effects on learning. If this null effect is observed, however, we would not know if it was due to equivalent learning or a flaw in the experimental method (e.g., low power). Thus, we needed to include a third condition to ensure that the method was working properly. For the third condition, we chose to have individuals solve problems while observing worked examples of the same problems being solved by an expert. Many studies have demonstrated high learning gains when individuals study worked examples and solve problems in various combinations (Atkinson et al., 2000; Renkl, 2005). Because Chi et al. (2008) found that learning of pairs observing tutoring matched that of actual tutoring, and as it is widely believed that tutoring is more effective than individual study of examples and problem solving (VanLehn, 2009), we expected that our target group, that is, pairs observing tutoring,

would gain more than a control group consisting of individuals observing worked examples.

The Current Study

The LearnLabs (www.learnlab.org) of the Pittsburgh Science of Learning Center (PSLC) acted as facilitators for the investigation, bringing researchers together with schools and teachers to scientifically test learning theories in classrooms. The current study was conducted in the PSLC's Physics LearnLab. This LearnLab consists of introductory physics courses at the United States Naval Academy. These courses use the Andes system (VanLehn et al., 2005) provided by the PSLC as the homework portion of their course.

The Andes system provides introductory college-level physics homework problems. The Andes program is not a complete instructional system but rather a coach that helps the student solve homework problems. It plays the same role in the course as a workbook, except that it provides immediate feedback and hints while students are solving problems. It encourages certain problem-solving practices (e.g., drawing vectors instead of imagining them) to increase conceptual understanding of physics. The problem solving involves algebra, trigonometry, and vectors, but not calculus. In this way, it is intended to be used with almost any course's textbook, lectures, and labs. The system tracks the student's progress and provides him or her with a score based on the student's problem solving for each problem. As previous research on vicarious learning has shown the Andes system to be able to promote both procedural learning (Fox Tree, 1999) and deeper conceptual learning (Chi et al., 2008; Craig et al., 2006), it provides an ideal bridge for moving into the classroom. Andes is freely available on the Internet.[1]

In the study reported here, we evaluated collaboratively observing tutoring in the classroom. In doing so, we compared collaborative observers of tutoring videos during problem solving in Andes (collaboratively observing tutoring condition) against two control conditions. The first control condition required pairs of students to collaboratively observe a worked example video during problem solving in Andes (collaboratively observing examples condition). In the second control condition, individually observing examples, individual students viewed worked example videos alone while problem solving in Andes. Because the Andes system provides video explanations for the learners on select problems, this control was analogous to the help that was normally provided to the student in the course. Neither Chi et al. (2008) nor Craig et al. (2004, 2006) found learning gains for individuals who observed tutoring when compared with various controls. Therefore, the condition in which individuals observed tutoring was not taken into the classroom so as to avoid exposing students to an ineffectual learning condition.

Two contrasting hypotheses were tested in this design. The *active observing hypothesis* predicted that the learners in the collaboratively observing tutoring condition would outperform those in other conditions because of the highly dynamic tutoring session. Thus, the tutoring videos contained dialogue features (e.g., turn taking, pauses, and affect) and expert tutoring elements (e.g., corrections and scaffolding) designed to promote more active engagement with the video material. In contrast, the passive information display from the worked examples did not include such features. This hypothesis generated prediction of the following pattern of learning gains:

Collaboratively observing tutoring

> collaboratively observing worked examples

= individually observing worked examples. (1)

An alternative hypothesis, the *content equivalency hypothesis*, is based on the premise that the content is what really matters. If learners receive equivalent content, the method in which the material is presented should not influence learning (Klahr & Nigam, 2004). As all participants in our study were exposed to the same content, this hypothesis predicted the following pattern of learning gains:

Collaboratively observing tutoring

= collaboratively observing worked examples

= individually observing worked examples. (2)

Method

Participants

United States Naval Academy (USNA) students (ages 18–19 years; $N = 67$) from three sections of the PSLC Physics LearnLab participated in this study. Participation in the study was a mandatory learning experience integrated into the laboratory section of the class, but students' data were used in the study only with their consent. Just one student did not give consent. This resulted in an n of 10 for the individually observing examples condition, an n of 26 for the collaboratively observing examples condition, and an n of 30 for the collaboratively observing tutoring condition. Four participants did not complete any homework problems and were excluded from the long-term assessments. This left an n of 9 in the individually observing examples condition, an n of 25 in the collaboratively observing examples condition, and an n of 28 in the collaboratively observing tutoring condition.

Because of the nature of classroom research, the number of participants tends to be a fixed small number, which can lead to statistical power problems. In the original Chi et al. (2008) data, the observed effect size of collaboratively observing tutoring was large ($d = .97$) when compared with individually observing tutoring controls. A power analysis (Cohen, 1988) with the large effect size indicated the need for a total of 60 participants to reach a standard $\beta = .80$ level for power with $\alpha = .05$. On the basis of this power analysis conducted with the G*Power system (Faul, 2008), the limited sample size was deemed to be sufficient.

Materials

Andes tutoring system. The Andes tutoring system (VanLehn et al., 2005) provides introductory college level physics homework problems (see Figure 1). The system was selected for use in this study because it was integrated into the Physics LearnLab sections of the USNA's introductory physics courses as the homework for

[1] See http://www.andestutor.org/ for further details on the Andes system.

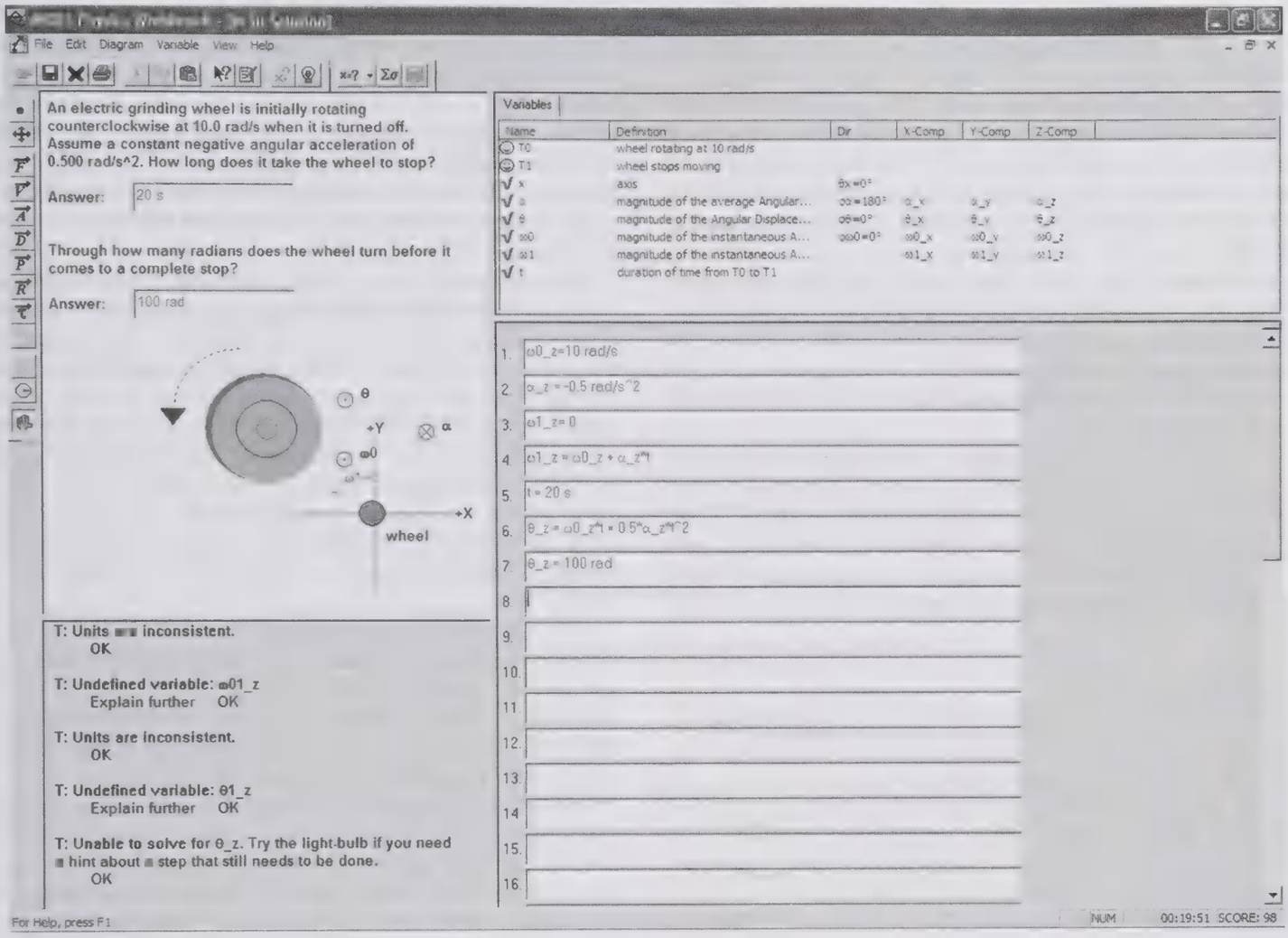


Figure 1. Screen shot of the Andes Physics Tutoring System with one of the two training problems on rotational kinematics.

the course. In addition to homework, the system was implemented as both the context of the learning videos and the problem-solving domain for training. However, because the study investigated the effect of observing videos, the help and feedback functions were disabled during the in-class training. The fully functioning version of Andes was available to students while they solved the course homework and on immediate post-training assessments.

Learning materials. Two sets of videos were observed by the learners. The videos were informationally equivalent in that they covered the same problem-solving steps in the same order and gave the same conceptual information. The TechSmith Camtasia® studio software package (TechSmith, 2006) was used to capture and edit all videos.

One set of videos consisted of an expert working out solutions for two Andes problems on rotational kinematics. There were two videos, one for each problem. In these videos, the expert, a retired USNA professor of physics with a PhD in the subject, solved two Andes problems while verbally presenting the relevant conceptual knowledge for each step. The videos presented the actions performed on the screen along with the expert's voice. In these videos, the expert covered the same steps in the same order as were covered in the tutoring session described below. Both sets of videos were approximately 22 min long.

The second set of videos consisted of recordings of an expert tutoring session on rotational kinematics. For these videos, the same physics expert tutored an intermediate-level tutee who had completed an introductory physics course that included rotational kinematics but did not have a degree or advanced training in physics. Again there were two videos, one for each problem. The videos were recorded by the same method as the worked example and displayed screen activity with voices of both the tutor and the tutee. The tutoring videos for the experiment were selected from a pool of five tutoring sessions on the basis of voice quality, the tutee's pretest being above 50% correct, the posttest score being above 90% correct, and complete topic coverage.

Immediate learning measures. Two isomorphic multiple-choice tests were used as pretest and immediate posttests. Both tests consisted of 12 four-choice questions assessing conceptual knowledge of rotational kinematics. The tests were counterbalanced across participants to prevent order effects.

In addition, three Andes problems were used as immediate posttest competency measures. The problems were designed as near-transfer problems using the same knowledge as in the training problems (see Appendix). The help and feedback features of the Andes system were available while participants completed the near-transfer problems. The Andes scoring rubric (VanLehn et al.,

2005) subtracted points for errors and overuse of help. Students were familiar with the scoring rubric as it was used for their homework.

Long-term retention and transfer. The long-term measures consisted of the students' Andes homework scores. They were instructed to complete these homework problems at any point between when they were assigned and the section test. Assignment of the homework problems occurred after the training session was completed.

The homework problems were divided into three categories. There was one long-term retention problem, three long-term near-transfer problems, and three long-term far-transfer problems. The long-term retention problem was one of the two problems taught during training. The homework problems listed by category can be found in the Appendix.

Equipment. During the training phase, participants shared a laptop computer with two Belkin headphone splitters that allowed for two headsets and microphones to be used on the same machine by the 2 participants. Participants' on-screen problem-solving activity and verbal interactions during training were captured using the TechSmith Camtasia recorder (TechSmith, 2006).

Design and Procedure

The current in vivo study implemented a pretest–posttest design with a long-term classroom impact measure to determine differences among the three experimental groups: collaboratively observing tutoring condition, collaboratively observing examples condition, and individually observing examples condition. The USNA students were randomly assigned to one of three conditions by lab tables. Thus, whereas they were allowed to work with their normal lab partner, they were blind to the condition of their lab table until after they selected their table. Following assignment of conditions, participants provided informed consent; any questions about the procedure were answered by the experimenter. Because this was a classroom setting with the instructor present, participants were given the option of contacting the experimenter outside of class if they wished to be excluded from the study.

After the informed consent process, all participants watched a brief 4-min video. This video introduced them to the terms and basic concepts of rotation to ensure that they had the prerequisite knowledge for completion of their problem-solving task. Afterward, all participants individually completed the multiple-choice pretest.

The learners then performed the training task, which consisted of completing two Andes problems with the aid of video solutions for each problem. They solved rotational kinematics problems using the Andes tutoring system while simultaneously watching either the tutoring session or the worked example videos showing the same problems being solved. Their voice and onscreen problem-solving activity were recorded.

All participants worked at their own pace on this task. However, there was no significant difference in the amount of time groups worked on the training problems, $F(2, 60) = 0.53$, $p = .59$ (for individually observing worked examples, $M = 32$ min; for collaboratively observing worked examples, $M = 28$ min; and for collaboratively observing tutoring, $M = 29$ min).

The students were assessed individually immediately after training with a multiple-choice posttest and three immediate near-transfer problems performed in Andes on rotational kinematics. Once the participants indicated they were done with the training session, each participant was given a multiple-choice posttest, which he or she completed individually. After the participant finished the multiple-choice test, the experimenter administered the Andes problems, which each participant also completed individually. Participants were given as much time as they needed to complete these tasks. All participants finished before the end of the class period.

Andes homework data. Long-term measures consisted of Andes homework problems that students completed in an unsupervised setting (their dorm rooms, typically). On average, the students completed their homework 26 days after the training session; there were no significant differences for time delays among groups. The 26-day delay might seem excessive, but it was in line with participants' completion of other physics homework problems. This delay was due to a standard class deadline set by the instructor. Completion of homework for the current section being covered was not requested until the end of the section. This resulted in most homework being completed on a fairly delayed basis.

Although the instructor encouraged students to help each other, students were required to solve their own Andes problems. The instructor was adept at using log data to detect cases where one student copied another's homework, so this rarely happened. Log data were harvested from the PSLC DataShop, which routinely collects Andes log data as the students do their homework.

Students could access the hints available in the Andes program and were permitted to consult their textbook, their friends, and even the instructor as they completed their homework. The long-term measures served as a type of dynamic assessment (Bransford & Schwartz, 1999; Haywood & Tzuriel, 2002), measuring the students' ability to transfer their learning to authentic instructional situations rather than operating within the sequestered setting of standard tests.

Andes homework was scored by an automatic metric that gave participants credit for correct steps while solving the problem and deducting points for errors and help requests. A total of 100 possible points per problem could be earned. The participant's average Andes score was calculated for each category and reported with the average total possible score of 100. If a participant did not complete any homework problems, he or she was excluded from this data set.

Results and Discussion

Immediate Learning Measures

Multiple-choice data. A series of analyses of variance (ANOVAs) was performed on the learners' multiple-choice data from the immediate pretests and posttests. The ANOVA performed on pretest data did not reveal significant differences among groups. Whereas the multiple-choice test showed that all students gained significantly from pretest to posttest, $F(1, 65) = 14.99$, $p < .001$, $\eta^2 = .231$, there were no significant differences among conditions on learning gains with the multiple-choice data, $F(1, 63) = 0.13$, $p = .877$, $\eta^2 = .003$. That is, the collaborative observers of tutoring, the collaborative observers of worked examples, and the individual observers of worked examples all seem to have learned

the same amount according to the multiple-choice data. Means and standard deviations for pretest and posttest data for all conditions are given in Table 1.

Immediate near-transfer data. An ANOVA was performed on the average Andes score across the three Andes near-transfer problems that were given immediately after the training (see Table 1 for Andes score means and standard deviations). As with the multiple-choice data, there were no significant differences among conditions, $F(1, 63) = 0.25$, $p = .782$, $\eta^2 = .001$. This lack of significance among conditions in the immediate learning data is consistent with the content equivalence hypothesis.

Long-Term Learning Measures

Long-term retention data. An ANOVA was performed on the participants' long-term retention data to determine differences among groups. This analysis revealed a significant difference between conditions, $F(2, 59) = 3.44$, $p < .05$, $\eta^2 = .104$. We performed a priori orthogonal contrasts to test our predictions. These tests revealed that students in the collaboratively observing worked examples condition did not significantly differ from students in our individually observing worked examples condition on their long-term retention tests, $t(59) = 0.67$, $p = .503$. Students in the collaboratively observing tutoring condition were then compared with the combined participants of the individually observing worked examples condition and the collaboratively observing worked examples condition. This contrast was significant and in favor of participants in the collaboratively observing tutoring condition, $t(59) = 2.61$, $p < .05$, $d = .68$. See Table 2 for means, standard deviations, and standard errors for long-term retention data.

Additionally, the mean time to solve the retention problem was 548.39 s (approximately 9 min). This is significantly longer than the 293.29 s (approximately 5 min) per Andes physics problem during the immediate posttest, $F(1, 62) = 21.85$, $p < .001$, $h^2 = 0.358$. This significant time difference provides some confirmation for our interpretation that students took the long-term retention problem and other homework problems much more seriously than the immediate posttest problems in that participants spent more time attempting to solve each problem. However, this is only an assumption of ours given the data provided. It does not rule out possibilities that observing tutoring led students to perform other beneficial behaviors, such as seeking help from instructors or discussing problems with other students. Future work should include follow-up surveys or interviews to determine the specific mechanisms that produced this effect.

Long-term near-transfer data. We conducted an ANOVA on the participants' near-transfer data to determine differences among groups. This analysis revealed a significant effect of condition, $F(2, 59) = 4.39$, $p < .05$, $\eta^2 = .129$. We again performed a priori contrasts to test our predictions. These tests revealed that once again, in the near-transfer data, participants in the collaboratively observing worked examples condition were not significantly different from those in the individually observing worked examples condition, $t(59) = 0.21$, $p = .834$. The students' data from the collaboratively observing tutoring condition were then compared with the students' data of the combined individually observing worked examples and collaboratively observing worked examples conditions. This contrast was significant in favor of participants in the collaboratively observing tutoring condition, $t(59) = 2.85$, $p < .01$, $d = .74$. See Table 2 for means, standard deviations, and standard errors for the learners' near-transfer data.

Long-term far-transfer data. We conducted an ANOVA on the participants' far-transfer data to determine differences among groups. This analysis revealed a significant effect of condition, $F(2, 59) = 4.89$, $p < .05$, $\eta^2 = .142$. We conducted a priori contrasts to test our predictions. These tests revealed that participants in the collaboratively observing worked examples condition were not significantly different from our individually observing worked examples condition, $t(59) = 0.05$, $p = .963$. We then compared the students' data from the collaboratively observing tutoring condition with the data from the combined individually observing worked examples and the collaboratively observing worked examples conditions. This contrast was significant in favor of participants in the collaboratively observing tutoring condition, $t(59) = 2.96$, $p < .05$, $d = .77$. See Table 2 for means, standard deviations, and standard errors for far-transfer data.

As can be seen in Table 2, we observed a different pattern of data between our immediate learning measures and long-term learning measures. No group differences were observed in our immediate learning measures. In our long-term learning measures, collaboratively observing tutoring outperformed both individually and collaboratively observing examples. Although the results of our long-term data argue strongly in favor of our active observing hypothesis, the null effect on immediate assessment is consistent with the content equivalency hypothesis. The observed reversal in effects could be due to two possible factors. First, the students

Table 1
Means and Standard Deviations for Immediate Learning Measures for the Three Conditions

Condition	Multiple-choice pretest		Multiple-choice posttest		Gain scores		Immediate near transfer	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Individually observing worked examples	.58	.23	.69	.18	.11	.20	64	22
Collaboratively observing worked examples	.56	.22	.65	.20	.09	.20	61	29
Collaboratively observing tutoring	.58	.18	.66	.17	.08	.16	57	30

Note. Values shown are proportion correct on multiple-choice pretest problems, proportion correct on multiple-choice posttest problems, proportion gain scores (scores on multiple-choice posttest minus scores on multiple-choice pretest), and immediate near-transfer score (Andes problem-solving score; out of a possible 100).

Table 2
Means, Standard Deviations, and Standard Errors for Long-Term (Robust) Learning Measures From Andes Homework Scores Across the Three Conditions

Condition	Long-term retention			Long-term near transfer			Long-term far transfer		
	M	*SD*	*SE*	*M*	*SD*	*SE*	*M*	*SD*	*SE*
Individually observing worked examples	73	17	6	68	33	11	53	42	14
Collaboratively observing worked examples	78	22	4	70	30	6	54	31	6
Collaboratively observing tutoring	88	13	2	88	13	2	77	22	4

Note. Andes homework scores are out of a possible 100 points.

might have taken the long-term assessment more seriously than the immediate tests. This could be attributable to the homework being part of the students' course grade. Thus, we assume that because the students likely took this series of Andes problems more seriously, we were able to detect an effect for our manipulations.

Alternatively, the immediate assessments may tap only shallow knowledge, which is provided equally well by all three conditions. For instance, all three types of instruction might have allowed students to remember the problem-solving steps immediately afterwards and then to use them for a copy-and-edit style of problem solving (VanLehn, 1998) but not to gain the deeper understanding that is needed for long-term performance and transfer.

Analysis of Students' Behavior During the Process of Collaboratively Observing

In order to test the claims of the active learning hypothesis, we analyzed the students' problem-solving behavior during training to determine some potential causes of the differences between the two collaboratively observing conditions. The videos were analyzed at both a macro level and a micro level.

At the macro level, the pairs' interactions were coded on the basis of their task engagement levels. In these codings, two raters viewed 30-s excerpts from each pair's recorded session. These excerpts were taken at 10% intervals from each collaborative observer session (e.g. at 3 min, at 6 min, and so forth for a 30-min video). This resulted in a total of 10 selections per video, for a total of 450 codings conducted. Active engagement was coded when the collaborative pair was discussing the problem-solving task (planning), engaging in discussion to determine a discrepancy in a member's knowledge, engaging in explanation of a problem-solving step, or collaboratively engaged in the problem-solving task. A kappa score of .76 was obtained between the two coders. This kappa level was deemed a sufficient level of agreement (Cohen, 1960), and disagreements were worked out between raters. A *t* test conducted on the data from the engagement coding revealed that the pairs who collaboratively observed tutoring were more actively engaged in their problem-solving task than the pairs who collaboratively observed an example, $t(44) = 2.13$, $p < .05$, $d = 0.63$. See Table 3 for the mean proportion of time that collaborative learners in each group were actively engaged.

This pattern was replicated to a lesser extent in a microlevel analysis of each pair's problem-solving steps. Each step of the problem was coded on the basis of the way in which it was obtained by the collaborative pair. Steps could be copied directly from the video example or generated by the collaborative pair. Further, we examined whether the observing pairs were using the videos to help scaffold learning by coding whether the video was searched when attempting to solve a step. While the observed copying, $t(43) = 0.45$, *ns*, and generating, $t(43) = 0.51$, *ns*, behaviors were not significantly different across conditions, pairs observing tutoring did actively search the video more often when attempting to solve a step, $t(43) = 2.09$, $p < .05$. Pairs of learners observing tutoring were more likely to search the video to verify a step in the solution and not just to copy it. Thus, these learners were more likely to find discrepancies between their internal mental model and that presented in the video. This discrepancy detection is a key component of active learning (Chi, de Leew, Chiu, & LaVancher, 1994; Wittrock, 1989). See Table 3 for the means and standard deviations of the codings by condition.

Table 3
Means and Standard Deviations for Active Engagement Levels (Proportion) and Active Search as Well as Means and Standard Deviations for the Proportion of Problem-Solving Steps That Collaborative Observers Either Copied or Generated

Condition	Active engagement		Active search		Copied steps		Generated steps	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Collaboratively observing worked examples	.48	.19	.00	.00	.30	.32	.68	.30
Collaboratively observing tutoring	.60	.19	.09	.20	.27	.23	.72	.23

Note. Values for active engagement and active search, respectively, are based on proportion of time during collaboration and active searching of video during problem solving.

General Discussion

The overall pattern of data partially supports our active observing hypothesis. Although immediate assessments of competence showed no differences across conditions, students in the collaboratively observing tutoring condition outperformed the students in the other conditions for all three of our long-term learning measures. Further, as claimed by the active learning hypothesis, learners in the collaboratively observing tutoring group displayed more active learning processes, with both more active engagement and active searching of the material during training.

These results suggest that when students collaboratively observe tutoring, they tend to have more active collaboration, which is followed by increases on long-term learning measures. This finding provides more evidence that active learning can improve learning from observing (Chi et al., 2008; Gholson & Craig, 2006) and is consistent with the literature on self-explanation (Chi et al., 1989) and multimedia learning (Mayer, 2001; Wittrock, 1989) that indicates the importance of active learning for deeper conceptual learning or transfer of learning to occur. More important, this finding shows that collaboratively observing tutoring while problem solving is a useful tool for improving learning outcomes in classroom settings when compared with traditional worked examples.

The current research was conducted in physics problem solving. Future research is required to test the generalizability of this finding to other domains, to younger populations, and to classroom instruction. However, past research has shown that vicarious learning techniques with a scripted question-led dialogue between a virtual tutor and tutee is effective for teaching 8th- to 11th-grade students (Craig et al., 2008; Gholson et al., in press). Vicarious learning has also been shown to be effective in the domains of conceptual physics (Gholson et al., in press), the circulatory system (Craig et al., 2008), and computer literacy (Craig et al., 2006; Gholson et al., in press). It is feasible that collaboratively observing tutoring would be a viable learning method in these areas as well.

Given the consistent findings of our long-term data and analyses of training transcripts, the null results on our immediate measures may be a result of some students rushing through the assessments perhaps because these assessments, unlike the long-term ones, did not affect the course grades. However, this explanation is not the only one possible. For instance, collaboratively observing tutoring might lead to later changes in student study behavior. That is, watching the tutee on the video struggle but ultimately learn might encourage students to study harder themselves.

However, the difference of our immediate low stakes versus our long-term high stakes assessments in classrooms settings could be a warning to researchers in the Learning Sciences as they move laboratory research into the classroom. Students in the laboratory are volunteers that usually receive compensation for their participation in the form of money or course credit. This compensation is often proportional to the time they work. This may create a demand to take all of the tasks seriously. On the other hand, students who perform tasks as part of their normal instruction, even if they have consented to have their data used by experimenters, may apply their normal prioritization schemes. In our case, students might have felt that after they received the instruction offered that day (e.g., observing the videos while solving problems), they could finish learning rotational kinematics at home and thus might have viewed the posttesting as merely an untimely nuisance. However, follow-up experiments are needed to verify this claim.

Because observing tutoring involves pairs of students watching a tutoring video together while collaboratively solving problems, it is most easily deployed as a classroom activity provided that each pair has a computer or a video player that the students can control. However, collaborative viewing could also be useful as a homework activity if a pair of students can meet after school or can collaborate remotely. The success of tutoring observation also suggests the utility of taking another look at standard instructional practice, wherein the teacher tutors a student in front of the class.

The current study suggests that observing tutoring is an effective alternative to standard instructional methods such as studying worked examples. Other laboratory research suggests that tutoring observation is as effective as interacting with both a human tutor (Chi et al., 2008) and an intelligent tutoring system (Craig et al., 2006). This study is the first to test observing tutoring in vivo, that is, as part of normal class instruction. Although the classroom context appears to have affected our assessments immediately after training, the instruction itself seems to have survived the transition from laboratory to a real classroom while retaining its effectiveness. If these results continue to replicate in the classroom, then we would have an effective alternative to labor-intensive human tutoring and costly intelligent tutoring systems.

References

Alessi, S. M., & Trollip, S. R. (1991). *Computer-based instruction: Methods and development* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4,* 167–207.

Atkinson, K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70,* 181–214.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer based instruction. *Journal of Educational Computing Research, 13,* 109–125.

Bandura, A. (1969). *Principles of behavior modification.* New York: Holt, Rinehart, & Winston.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In *Review of research in education* (Vol. 24, pp. 61–100). Washington, DC: American Educational Research Association.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145–182.

Chi, M. T. H., de Leew, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439–477.

Chi, M. T. H., Roy, M., & Hausmann, R. G. M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science, 32,* 301–341.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprentice-

ship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Hillsdale, NJ: Erlbaum.

Cox, R., McKendree, J., Tobin, R., Lee, J., & Mayes, T. (1999). Vicarious learning from dialogue and discourse. *Instructional Science, 27,* 431–458.

Craig, S. D., Brittingham, J., Williams, J., Cheney, K. R., & Gholson, B. (2009). Incorporating vicarious learning environments with discourse scaffolds into physics classrooms. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. C. Graesser (Eds.), *Artificial intelligence in education, building learning systems that care: From knowledge representation to affective modeling* (pp. 680–682). Washington, DC: IOS Press.

Craig, S. D., Driscoll, D., & Gholson, B. (2004). Constructing knowledge from dialog in an intelligent tutoring system: Interactive learning, vicarious learning, and pedagogical agents. *Journal of Educational Multimedia and Hypermedia, 13,* 163–183.

Craig, S. D., Gholson, B., Ventura, M., Graesser, A. C., & the Tutoring Research Group. (2000). Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. *International Journal of Artificial Intelligence in Education, 11,* 242–253.

Craig, S. D., Graesser, A., Brittingham, J., Williams, J., Martindale, T., Williams, G., Gray, R., Darby, A., & Gholson, B. (2008). An implementation of vicarious learning environments in middle school classrooms. In K. McFerrin, R. Weber, R. Weber, R. Carlsen, & D. A. Willis (Eds.), *The Proceedings of the 19th International Conference for the Society for Information Technology & Teacher Education* (pp. 1060–1064). Chesapeake, VA: Association for the Advancement of Computing in Education.

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). Deep-level reasoning questions effect: The role of dialog and deep-level reasoning questions during vicarious learning. *Cognition and Instruction, 24,* 563–589.

Derry, S. J., & Potts, M. K. (1998). How tutors model students: A study of personal constructs in adaptive tutoring. *American Educational Research Journal, 35,* 65–99.

Driscoll, D., Craig, S. D., Gholson, B., Ventura, M., Hu, X., & Graesser, A. C. (2003). Vicarious learning: Effects of overhearing dialogue and monologue-like discourse in a virtual tutoring session. *Journal of Educational Computing Research, 29,* 431–450.

Faul, F. (2008). *G*Power* (Version 3.0.10). [Computer software]. Kiel, Germany: University of Kiel.

Fox Tree, J. E. (1999). Listening in on monologues and dialogue. *Discourse Processes, 27,* 35–53.

Gholson, B., & Craig, S. D. (2006). Promoting constructive activities that support vicarious learning during computer-based instruction. *Educational Psychology Review, 18,* 119–139.

Gholson, B., Witherspoon, A., Morgan, B., Brittingham, J., Coles, R., Graesser, A. C., et al. (in press). Exploring the deep-level reasoning effect among eighth to eleventh graders and college students in the domains of computer literacy and Newtonian physics. *Instructional Science.*

Graesser, A. C., Jeon, M., & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes, 45,* 298–322.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H., Ventura, M., Olney, A., et al. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers, 36,* 180–193.

Graesser, A. C., Person, N., Harter, D., & the Tutoring Research Group. (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education, 12,* 257–279.

Haywood, H. C., & Tzuriel, D. (2002). Applications and challenges in dynamic assessment. *Peabody Journal of Education, 77*(2), 40–63.

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70,* 151–179.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15,* 661–667.

Latham, G. P., & Saari, L. M. (1979). The importance of supportive relationships in goal setting. *Journal of Applied Psychology, 64,* 151–156.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* New York: Cambridge University Press.

Mayer, R. E. (2001). *Multimedia learning.* New York: Cambridge University Press.

Meltzoff, A. N. (2005). Imitation and other minds: The "like me" hypothesis. In S. Hurley & N. Chater (Eds.), *Perspectives on imitation: From cognitive neuroscience to social science* (pp. 55–77). Cambridge, MA: MIT Press.

Renkl, A. (2005). The worked-out-example principle in multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 229–246). Cambridge, United Kingdom: Cambridge University Press.

Rogoff, B., Paradise, R., Mejía Arauz, R., Correa-Chávez, M., & Angelillo, C. (2003). Firsthand learning through intent participation. *Annual Review of Psychology, 54,* 175–203.

Rosenthal, R. L., & Zimmerman, B. J. (1978). *Social learning and cognition.* New York: Academic Press.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21,* 211–232.

Shebilske, W., Jordan, J., Goettl, B., & Paulus, L. (1998). Observation versus hands-on practice of complex skills in dyadic, triadic, and tetradic training-teams. *Human Factors, 40,* 525–540.

Silliman, E. R., & Wilkinson, L. C. (1994). Discourse scaffolds for classroom intervention. In G. P. Wallach & K. G. Butler (Eds.), *Language learning disabilities in school-aged children and adolescents* (2nd ed., pp. 27–52). Boston: Allyn & Bacon.

TechSmith. (2006). *TechSmith Camtasia* (Version 3.1) [Computer software]. (2006). East Lansing, MI: Author. Available from www.techsmith.com

VanLehn, K. (1998). Analogy events: How examples are used during problem solving. *Cognitive Science, 22,* 347–388.

VanLehn, K. (2009). *The interaction plateau: Less interactive tutoring is often just as effective as highly interactive tutoring.* Manuscript submitted for publication.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., et al. (2005). The Andes Physics Tutoring System: Lessons learned. *International Journal of Artificial Intelligence and Education, 15,* 147–204.

Wittrock, M. C. (1989). Generative processes of comprehension. *Educational Psychologist, 24,* 345–376.

Appendix

Andes Tutoring System Training and Assessment Items

Andes Training Problems (Andes Label: KR1A, KR3B)

1. A wheel is rotating counterclockwise as a constant angular velocity of π rad/s. through what angle does the wheel rotate in 60.0 s?

2. An electric grinding wheel is initially rotating counterclockwise as 10.9 rad/s when it is turned off. Assume a constant negative angular acceleration of 0.500 rad/s^2. How long does it take the wheel to stop? Through how many radians does the wheel turn before it comes to a complete stop?

Andes Immediate Near-Transfer Problems (Andes Labels: KR1C, KR3C, KR4B)

1. A wheel rotates counterclockwise at a constant angular velocity of 2.5 rad/s. How long does it take the wheel to rotate through an angle of 210 rad?

2. The magnitude of the initial angular velocity of a wheel rotating counterclockwise is 30 rad/s. If the wheel takes 15 s to slow to a complete stop, what was the average angular acceleration of the wheel?

3. A wheel is initially at rest. If the wheel undergoes an average angular acceleration of 1.5 rad/s^2 over time, how fast would it be rotating 10 seconds later?
angular velocity:
After this time, the wheel continues to rotate at constant angular velocity.
What is the angular displacement of the wheel during a subsequent 20 second time interval?
angular displacement:

Andes Long-Term Retention Problem (Andes Label: KR1A)[A1]

1. A wheel is rotating counterclockwise at a constant angular velocity of π rad/s. through what angle does the wheel rotate in 60.0 s?

Andes Long-Term Near-Transfer Problems (Andes Labels: KR1B, KR2B, KR3A)

1. A wheel is rotating clockwise at a constant angular velocity of 3*π rad/s. What is the magnitude of the angular displacement of the wheel after 45.0 seconds?

2. The initial angular velocity of a wheel is π rad/s in a clockwise direction. If the wheel is speeding up with a constant angular acceleration of π/4 rad/s^2, what is the magnitude of the angular velocity of the wheel after 15.0 seconds?

3. The magnitude of the initial angular velocity of a wheel rotating counterclockwise is π rad/s. If the wheel is slowing down with an average angular acceleration of π/6 rad/s^2, how long does it take to stop?

Andes Long-Term Far-Transfer Problems (Andes Labels: KR4A, KR6A, and KR7A)

1. A wheel has an initial angular velocity of 3*π rad/s in a counterclockwise direction. If the wheel is slowing down with a constant angular acceleration of π/4 rad/s^2, through what angle does it turn before it reaches a final angular velocity of π rad/s in a clockwise direction?

2. Two fixed pulleys are attached by a fan belt. The radius of the first pulley is 0.030 m. The magnitude of its angular velocity is 2*π rad/s in a counterclockwise direction. If the radius of the second pulley is 0.020 m, what is the magnitude of its angular velocity if the fan belt does not slip?
Note: Consider rim1 and rim2 to be points on the rims of the pulleys in contact with the belt at the instant depicted. Use a relative position vector to represent the perpendicular distance of a rim point from the axis of rotation.

3. A wheel is rotating at a constant angular velocity of π rad/s in a clockwise direction. The radius of the wheel is 0.030 m. What is the magnitude of the linear velocity of a point halfway between the center of the axle and the outside edge of the wheel?
Note: use a relative position vector to represent the perpendicular distance of a rotating point from the axle.

Note. An asterisk within the Andes problems represents the multiplicative function.

[A1] Whereas Problem KR3B (along with Problem KR1A) was initially intended to be used as a long-tern retention problem, researchers had no control over the course content, and this problem was not assigned as homework by the instructor.

Received April 15, 2008
Revision received May 21, 2009
Accepted May 28, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 790–802

0022-0663/09/$12.00 DOI: 10.1037/a0016656

Practice Enables Successful Learning Under Minimal Guidance

Angela Brunstein, Shawn Betts, and John R. Anderson
Carnegie Mellon University

Two experiments were conducted, contrasting a minimally guided discovery condition with a variety of instructional conditions. College students interacted with a computer-based tutor that presented algebra-like problems in a novel graphical representation. Although the tutor provided no instruction in a discovery condition, it constrained the possible actions sufficiently that students could always discover the algebraic transformations they needed to learn. In Experiment 1, with ample practice for each new transformation, students performed better in the discovery condition than any instructional condition. In Experiment 2, with only a little practice for each transformation, students performed worst in the discovery condition. The authors suggest that the high levels of practice in the 1st experiment made students more efficient at discovering the algebraic transformations. When the cognitive demands were manageable, the discovery students may have more often encoded the algebraic transformations in mathematically correct ways.

Keywords: discovery learning, intelligent tutors, practice, cognitive load

There is a long history of advocacy of discovery learning that includes such intellectual giants as Rousseau, Dewey, and Piaget. Bruner (1961) is frequently credited as the source for the modern research on discovery learning in the last 50 years. Discovery learning is typically contrasted with direct instruction, and the contrast between the two is best conceived of as a continuum. At one end of the continuum, students are directly told what they are to learn; at the other end, students are left to find out what they are to learn through exploration. However, no learning experience is pure; students given direct instruction often find themselves struggling to discover what the teacher means, and all discovery situations involve some minimal amount of guidance, if only to tell the students to try to make sense of the situation. Moreover, the space of instructional strategies is hardly one-dimensional, and strategies that tend to the discovery end can vary substantially. Kirschner, Sweller, and Clark (2006) have used the term *minimally guided instruction* to refer to strategies that tend to this end of the spectrum.

Angela Brunstein, Shawn Betts, and John R. Anderson, Psychology Department, Carnegie Mellon University.

Angela Brunstein is now at the Department of Social and Decision Sciences, Carnegie Mellon University.

This research was supported by National Science Foundation Award REC-0087396 and Grant AFOSR–FA9550-07-1-0359 from the Defense Advanced Research Projects Agency. Angela Brunstein was partly supported by a Feodor-Lynen Research Prize from the Alexander von Humboldt Foundation.

An extended version of this article reporting an ACT-R model and its fit to the data is available from the ACT-R website, http://act-r.psy.cmu.edu/publications/index.php (Brunstein, Betts, & Anderson [2009]. *When minimal guidance does and does not work: Drill and kill makes discovery learning a success.* Unpublished manuscript)

Correspondence concerning this article should be addressed to John R. Anderson, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: ja@cmu.edu

Although minimally guided instruction continues to have its advocates (e.g., Fuson et al., 1997; Hiebert et al., 1996; Kamii & Dominick, 1998; von Glasersfeld, 1995), evidence and argument have been accumulating against it (e.g., Kirschner et al., 2006; Klahr & Nigam, 2004; Mayer, 2004; Rittle-Johnson, 2006). Indeed, in two of the responses to the Kirschner et al. (2006) criticisms of minimally guided learning, the authors (Hmelo-Silver, Duncan, & Chinn, 2007; Schmidt, Loyens, van Gog, & Paas, 2007) did not question the claim that minimally guided learning was bad. Rather, they questioned whether Kirschner et al. had it right in classifying problem-based and inquiry learning as minimally guided.

A distinction that one frequently finds in the literature (e.g., Baroody, Lai, & Mix, 2006; Mayer, 2004; Shulman & Keisler, 1966) is between *pure discovery* and *guided discovery*. In typical guided discovery, the teacher provides "hints, direction, coaching, feedback and/or modeling" to keep the student on track, whereas in pure discovery the teacher provides "little or no guidance" (Mayer, 2004, p. 15). Mayer argued that pure discovery is almost always worse than direct instruction because students often fail to come in contact with the material to be learned. On the other hand, he argued that guided discovery can be more successful than direct instruction because it leads to integration of the new information with existing information.

It is ambiguous whether what we call a discovery condition in this article should be called guided discovery. In this study, students interacted with a computer-based tutoring system. In the discovery condition, the tutor does not provide any hints, direction, coaching, or modeling, as described by Mayer (2004). However, the nature of the computer interface means that it does provide feedback, sometimes more immediate and sometimes more delayed, on whether students have performed correct or incorrect actions. This feedback can be viewed as providing "hotter" or "colder" evaluations of student actions at some delay from these actions. The computer interface also limits the search space that the students have to explore in trying to make their discoveries.

Perhaps it would be most accurate to characterize our discovery[1] condition as minimally guided discovery, in line with the usage of Kirschner et al. (2006).

This research is part of an effort to understand the contribution of instructional content in cognitive tutors that are based on cognitive models of how students solve mathematics problems. Cognitive tutors have been shown to have some success in the teaching of mathematics (Anderson, Corbett, Koedinger, & Pelletier, 1995; Koedinger, Anderson, Hadley, & Mark, 1997). They are deployed in over 2,600 schools throughout the United States and interact with approximately 500,000 students each year (Koedinger & Corbett, 2006; Ritter, Anderson, Koedinger, & Corbett, 2007; Ritter, Haverty, Koedinger, Hadley, & Corbett, 2008). It has been noted that students working with the tutors can become too dependent on the help they receive, and there has been extensive research on how to make these tutors more effective (e.g., Aleven, McLaren, Roll, & Koedinger, 2006; Baker et al., 2008; Heffernan, Koedinger, & Razzaq, 2008). The current research investigates whether one can get better learning by removing some of the guidance that the tutor provides on how to solve problems. More specifically, this research looks at how degree of practice can influence the effectiveness of a minimally guided discovery condition. In this research, the instructional conditions we use mainly serve as reference points for understanding the effect of such manipulations on the discovery condition.

The mathematics topics taught by the tutors have a sufficiently rich combinatorial structure that it is not possible to provide students with direct instruction on all possible cases. Students must generalize what they learn on specific cases to new cases. For instance, in this research, after students learned to rewrite $(4 + x) + 3$ as $7 + x$, they were given the new problem $(5 + x) - 3$. Although the majority of students correctly generalized and rewrote this as $2 + x$, a significant minority displayed the error of rewriting it as $2 - x$. Making the correct generalization to this case can be viewed as minidiscovery informed by knowledge of the constraints of algebra and arithmetic. Thus, even though students are taught to rewrite $(4 + x) + 3$ as $7 + x$, and even though they know enough about subtraction and addition to extend that knowledge, they must still determine how to integrate that knowledge in the case of $(5 + x) - 3$. This is basically a minidiscovery. Students are better able to make such generalizations if they integrate what they are learning with their general knowledge of arithmetic. Mayer (2004) argued that this is more likely to happen with guided discovery, and we show that this can happen in the discovery environment we have created.

Figure 1 shows some screen images involving equation solving in the Carnegie Learning Cognitive Tutor (2007). In terms of the interface interactions, these are the simplest parts of the algebra curriculum, but they reflect the general model of interaction with the Cognitive Tutor. In Figure 1a, the student is presented with the equation $8y + (-6y) + 9 = 10$, and the student selects an operation to perform from a pull-down menu—in this case, the student has erroneously selected *Distribute*. The student then will receive feedback and eventually choose the correct operation of *Add/Subtract Terms*. When this correct operation is chosen, the tutor presents a display like Figure 1b, where the student must indicate the result of adding like terms by filling in a series of boxes. The resulting equation is represented in Figure 1c, and the student must choose a correct operation again. Upon doing so, the tutor once again presents a series of boxes (Figure 1d) where the student must indicate the terms being subtracted. This illustrates the basic cycle in the tutor in which the student selects some operation to perform (Figures 1a and 1c) and then executes the result of that operation (Figures 1b and 1d) by filling in some boxes. By isolating the individual operations and executions, the tutor is able to identify specific difficulties that the student is having and to provide instruction on those aspects.

The research reported here involves some extreme experimental manipulations that might well fail to result in mastery of algebra. Therefore, we did not want to study children learning linear algebra, lest our experimental manipulation hinder their future ability to master algebra. Rather, we developed a data-flow isomorph of linear algebra equations suitable for teaching to college students who have already mastered linear algebra. Essentially, because of the novel format, college students go through the process of learning to solve equations anew. If students fail to learn, as they did in some of the conditions reported here, it is at no loss to their competence with normal algebra. Nevertheless, the semantics that underlie operations in the data-flow isomorph are the same as in standard algebra. Therefore, learning to interact with the tutor corresponds with grasping the semantics of algebra to exploit its combinatorial structure. Figure 2 shows two examples of data-flow graphs that correspond to particular linear equations. Figure 2a is the isomorph of the equation $5x + 4 = 39$, and Figure 2b is the isomorph of the equation $(2x - 5x) + 13 + 9x = 67$. In such a diagram, a number comes in the top box and flows through a set of arithmetic operations; the result is the number that appears in the bottom box. Students are taught a set of graph transformations isomorphic to the transformations on the linear equations that result in simplifying the diagram. In the case of problems like those in Figure 2, these simplifications result in determining the value that has to go into the top box to produce the value in the bottom box. This is the equivalent to solving for a variable (i.e., x). However, some diagrams are the equivalent of expressions to be simplified (not equations to be solved), and their simplification requires the equivalent of algebra's collection of like terms and distribution of multiplication over addition. Anderson (2007) reported a behavioral comparison of children working with linear equations and adults working with the data-flow tutor. Although children were a bit more error prone, they learned and behaved very similarly.

We used the tutoring system described in Anderson (2007), which has the same interaction style as the Cognitive Tutor (2007) for algebra. It involved selecting parts of the diagram with a mouse, selecting transformations of those diagrams, and typing in the contents of the changed portions of the diagram. As with the Cognitive Tutor, this experimental tutor provides some initial guiding instruction, with further instruction if the student requests help or makes errors. Almost all of this instruction and guidance was removed in the discovery condition. We specify more about the interaction style and the various instructional conditions below in the description of Experiment 1.

[1] *Discovery* is intended to denote the name for a condition rather than an assertion about the true nature of that condition.

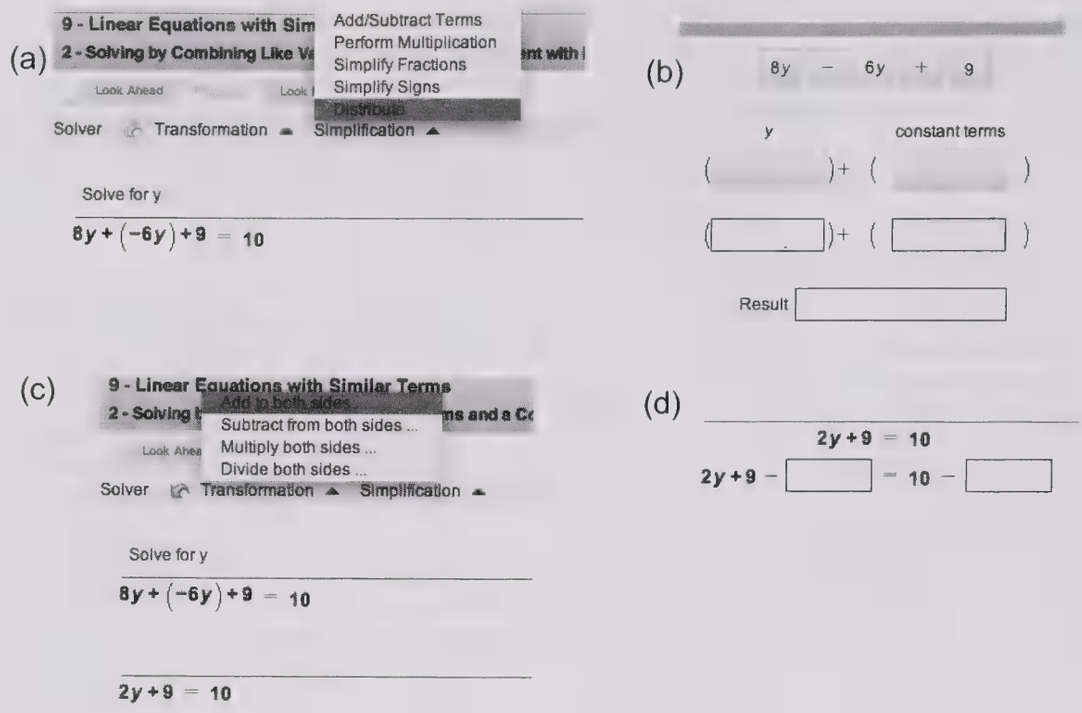


Figure 1. A representation of four steps in the solution of an equation with the algebra tutor: (a) selection of a transformation; (b) filling in of the transformation; (c) selection of an evaluation; (d) filling in of the evaluation.

Experiment 1

Much of the instruction provided by cognitive tutors comes in response to help requests or errors. Typical of most instruction in mathematics, this instruction involves a mix of verbal directions and pieces of worked examples. The first experiment reported here was an attempt to assess separately the contributions of the instruction and worked example and whether we could get better learning if we removed one or both. There was a *verbal direction* condition in which participants received abstract verbal instruction without any specific directions about how to solve a problem and a *direct demonstration* condition in which participants were told what to do in a specific case without receiving any general characterization of the action. To complete a factorial design, we crossed the use of verbal direction with direct demonstration. This created a *both* condition that was similar to the original condition of Anderson (2007), where students received both an abstract characterization and a demonstration of what to do. This also created the discovery condition, where there was no instruction accompanying the steps. Many experiments have compared examples, instructions, and a combination of the two (e.g., Charney, Reder, & Kusbit, 1990; Cheng, Holyoak, Nisbett, & Oliver, 1986; Fong, Krantz, & Nisbett, 1986; Reed & Bolstad, 1991), but these experiments have tended not to look at situations in which the participants receive no direction, as in our discovery condition. These experiments have produced somewhat different estimates of the relative contributions of examples and instruction, presumably reflecting differences in the material.

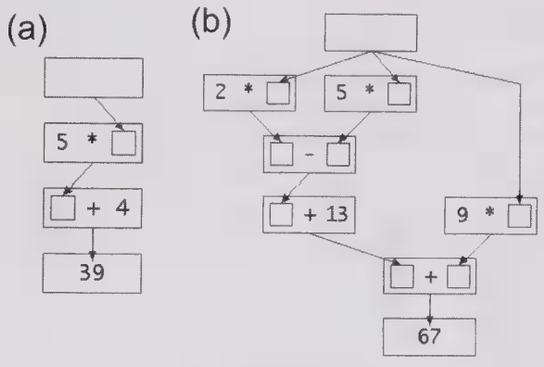


Figure 2. The data flow equivalents of (a) $5x + 4 = 39$ and (b) $(2x - 5x) + 13 + 9x = 67$.

Figure 3 illustrates the simple interface of the tutor. There are three interactions that students can have with the tutor. They can select boxes in the data-flow graph to operate on, select operations from the buttons to the right, and type values of the expressions into dialog boxes like the one illustrated in Figure 3. The correct combination of these actions can succeed in simplifying the diagram. This is much like the Carnegie Learning algebra tutor (Cognitive Tutor, 2007) in Figure 1.

Figure 4 uses a problem concerned with collection of like terms to illustrate the basic cycle that occurs throughout the curriculum. The problem in Figure 4 is the data-flow equivalent of $3 + (2x + 7)$. The first row in Figure 4 shows steps in the transformation of the problem from its original form to the equivalent of $(7 + 3) + 2x$; the second row shows steps in transforming this to the equivalent of $10 + 2x$. As the curriculum progresses, the problems become more complex and require more varied transformations, but their solutions always have the interaction cycle illustrated in Figure 4, as follows:

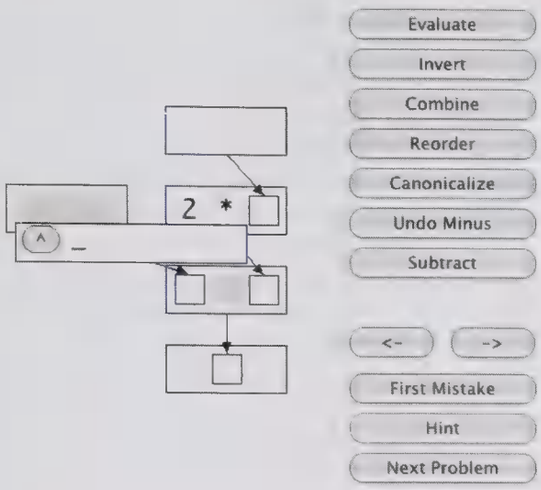


Figure 3. The tutor interface at the point where a result is to be typed.

1. The diagram begins in some neutral display (panels a and d), and the student selects some boxes to operate on. Later problems can require selection of as many as five boxes, and there can be a number of alternative correct choices about which sets of boxes to operate on next.

2. The selected boxes are highlighted in red (panels b and e), and the student selects some operation by clicking a button to the right of the diagram.

3. If a correct set of boxes and operations have been chosen, the diagram is transformed with a number of green boxes (panels c and f) to be filled in. The student can click these green boxes; an input dialog appears (see Figure 3), and the student can type information into the input dialog.

4. When the boxes are filled in with syntactically correct expressions (not necessarily the correct values), the diagram returns to a neutral state (panels d and g), ready for the next selection of some set of boxes.

The tutor's color conventions, as illustrated in Figure 4, are that red boxes indicate parts of the diagram selected for an operation and green boxes indicate information to be filled in. When the transformations are complete, the student clicks the *Next Problem* button. If the transformations have been correctly performed, the student can go onto the next problem. If there was an error, the student is informed that he or she can not go on to the next problem but has to correct the error. The *First Mistake* button takes the student to the state of the diagram before the first mistake. The arrow buttons allow students to move backward or forward by a single transformation.

The material used in this experiment comes from 12 sections over four chapters in *Algebra I* (Foerster, 1990), an algebra textbook that covers what is needed to solve linear equations. The first one or two problems in each section were used for instructing the material in that section. For these problems, instruction was volunteered whereas instruction was available on request for later problems (except for the discovery condition, where there was never any instruction).

The problem in Figure 4 was used for initial instruction in section 2.6 on combining constants. Table 1 shows the instruction that accompanied this problem. There is some general initial instruction and then instruction that accompanies each state of the problem. The instructional manipulations involved the state-by-

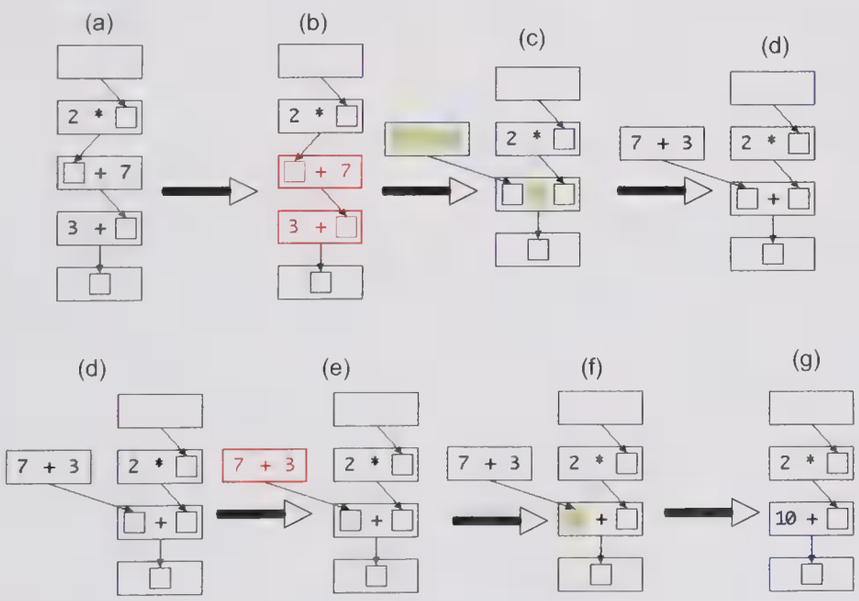


Figure 4. The steps in the solution of a combine problem, the data flow equivalent of $(2x + 7) + 3$, from section 2.6. Each picture is a different state of the diagram on its way to its simplest form. The two lines reflect the two transformations. The first line starts with the problem (a), then a part of the graph is selected and highlighted in red (b), then the combine operation is selected (resulting in c), and the parts are filled in (resulting in d). The second line starts with d from the previous line; then a part of the graph is selected for evaluation (e); the evaluation operation is selected (resulting in f); and the value is filled in (resulting in g).

Table 1
Instructions for Section 2.6 on Collecting Constant Terms

Initial general instructions	*One can collapse two boxes with + or − into a single box and preserve the value of the diagram. One can do the same thing with two boxes with * or /.*	
	Verbal directions	Direct demonstration
State		
a	*Find two boxes with addition or subtraction and click them*	*Click This* arrow
b	*Click the button labeled* Combine.	*Click This* arrow
c	*Click the little green box.*	*Click This* arrow
	Enter the operator from the box above.	*Type* + arrow
	Click the green big box.	*Click This* arrow
	Enter the number from the box above, then the operator from the box below, and then the number from the box below.	*Type 7 + 3* arrow
d	*Find a box with two numbers and an operator and click it.*	*Click This* arrow
e	*Click the button labeled* Evaluate.	*Click This* arrow
f	*Click the little green box.*	*Click This* arrow
	Find the answer by evaluating the box above and enter it.	*Type 10* arrow
g	*Your answer is correct. Type the* Next Problem *button.*	*Click This* arrow

state instruction. In the verbal direction condition, participants would receive instructions such as *Find two boxes with addition or subtraction and click them*, which provided guidance on how to perform the operation on this and similar problems without saying exactly what to do. In the direct demonstration condition, participants were told what to do in this specific case without stating any general characterization of the action. For instance, arrows would point to the two boxes with the instruction *Click this*. In the *both* condition, participants saw both forms of instruction, whereas in the discovery condition they saw neither. For section 2.6 on combining constants, the most critical transformation is between states like Figures 4c and 4d, where the participant must specify the content of the boxes in a way that preserves the value of the graph structure.

Participants in the discovery condition received none of the guidance illustrated for states (a) through (g) in Table 1, although they did see the initial general instructions. They had to try various actions and learn from the consequences of their actions. The following is a list of the sorts of errors that could be made and the feedback that would occur—this feedback was also available in the other conditions. The cases are listed in order of increasing delay of feedback.

1. Interface errors: If the student tried some action that the interface was not prepared to process (such as typing a number when there is no dialog box, as in Figure 3), the tutor did not change state. This lack of response was an immediate indication that the action had been rejected.

2. Operator errors: After students selected some boxes and an operator, they either saw the screen transform into a state with green boxes to be completed, indicating success, or saw an error message saying the operator would not apply to the boxes selected. The error message provided no explanation of why. It just provided the student with information that there was something wrong—either with the boxes they had selected or with the operator they had just selected.

3. Transformation errors: After selecting the operation, the student specified the transformation by typing material into the green boxes. Syntactically incorrect material was not accepted, but well-formed incorrect answers were accepted. Feedback on such semantic errors only occurred when the students got to the end of the problem and chose the *Next Problem* operator. If all the transformations had been correctly entered, the tutor went on to the next problem. Otherwise, it would give an error message: *Your answer is incorrect. Use the buttons (or the left and right arrow keys) and the* First Mistake *button to review your work and correct the mistake.* Delaying feedback on these errors to the end of the problem is analogous to what happens in algebra texts, where a student performs a series of transformations and can only check the final result against the answer in the back of the book. One exception to delay of feedback on transformation errors was on the first one or two instructional problems: If an incorrect result was typed into a green box, it was rejected with the error message *Your answer is incorrect.* The other possible exception is that students could choose to hit the *First Mistake* button to find out if they had made any transformation errors so far in their solution, but they seldom used this option.

These features had been built into the interface prior to designing our discovery condition. They were simply what remained after we removed the direct instruction to create the discovery condition. It might seem surprising that students could always discover how to solve the problems, but the interface limited the options enough that all students in the discovery condition eventually found solutions. Thus, the discovery students could be viewed as searching through a maze of interface actions, with the interface being responsive and restrictive enough that they eventually found their way out of the maze (i.e., solved the problem). The research tells us whether they actually learned anything about the domain from their search through this maze of actions.

Method

Materials

Participants solved 174 data-flow problems based on problems from 12 sections in the Foerster (1990) text that spanned the first four chapters. The first session took on average about 1 hr, whereas the second and third sessions took on average approximately 1.5 hr. Solving these problems required performing at least 674 operations. Below are the 12 sections and examples of the problems in their linear algebra equivalent form (the sections are labeled with chapter number first, followed by the section number within the chapter).

Section 1.1: Evaluating diagrams (14 problems). Teaches students how to evaluate the contents of boxes in the data-flow diagrams—for example, rewrite $(9 - 4) \times 2$ as 5×2, and rewrite this as 10.

Section 1.2: Input boxes (nine problems). Teaches students to evaluate a diagram given a value for an input box—for example, rewrite $(24/x) - 1$ and $x = 12$ as $24/12 - 1$, and this as $2 - 1$, and this as 1.

Section 1.7: Finding input values (25 problems). Teaches students to find the input values given single operations—for example, rewrite $x + 3 = 8$ as $x = 8 - 3$, and this as $x = 5$.

Section 2.6: Combining operations (20 problems). Teaches students how to combine constant terms—for example, rewrite $(5 + x) - 3$ as $(5 - 3) + x$, and this as $2 + x$.

Section 2.7: More on finding input values (16 problems). Teaches students to find the input values given two operations—for example, $2x + 3 = 19$—and to deal with asymmetric operators—for example, rewrite $10 - x = 2$ as $x = 10 - 2$, and this as $x = 8$.

Section 3.1: Reordering operations (six problems). Teaches students the graph equivalent of distribution—for example, rewrite $5 \times (x + 2) + 9$ as $[5x + (5 \times 2)] + 9$, and this as $(5x + 10) + 9$, and this as $(10 + 9) + 5x$, and finally as $19 + 5x$.

Section 3.2: Reordering and subtraction (nine problems). Teaches students to use reordering with subtraction in problems such as $9 - 2 \times (x - 4)$.

Section 3.4: Combining multiple input boxes (13 problems). Teaches students the equivalent of collecting variable terms—for example, rewrite $7x + 5x$ as $(7 + 5) \times x$, then as $12x$, and rewrite $5x + (6 - 2x)$ as $6 + (5 - 2) \times x$, then as $6 + 3x$.

Section 3.5: More on combining input boxes (12 problems). Deals with special cases like $2x + x$ (no coefficient before the variable) and $(6x + 3) - (6 - 2x)$ (combining both variables and constants).

Section 4.1: Finding input values in more complex problems (11 problems). Puts the operations together, building up to equations like $[(3x + 4) + 5x] + 6x = 32$.

Section 4.2: Finding input values in harder problems (21 problems). Builds up to equations like $3 \times (2x - 1) + 2 \times (x + 5) = 55$.

Section 4.3: Finding input values when two data-flow diagrams are equal (18 problems). Presents equations like $3x + 55 = 8x$.

The other sections of the Foerster (1990) textbook did not involve material relevant to linear equations. For instance, sections 2.1 to 2.5 were a review of signed arithmetic.

Participants and Conditions

Forty Carnegie Mellon undergraduates (23 male and 17 female; $M = 23$ years, $SD = 1.6$ years) took part in this study. They reported relatively high grades in their high school algebra courses (24 *A*s, 8 *B*s, 4 *C*s, 4 missing). Students participated in three single sessions, each lasting between 1 and 2.5 hr. In the first session, they went through the sections above from chapters 1 and 2; in the second session, they went through chapter 3; and in the third session, they went through chapter 4. They received performance-based feedback in the form of $ 0.07 per correctly performed operation in the tutor or a guaranteed minimum of $5 per half hour, whichever was greater. Fourteen students received performance-based pay in Session 1, none in Session 2, and 23 in Session 3.

Ten participants were randomly assigned to each of four conditions. The four conditions were defined by different combinations of instructions such as those in Table 1. The verbal direction condition received only the verbal directions; the direct demonstration condition received only the direct demonstration; the both condition received both; and the discovery condition received neither.

Measurements

The first problem in each section involved different combinations of the guided instruction in Table 1 (including none for the discovery condition). For sections 2.7, 3.4, and 3.5, the second problem in a section also involved guided instruction. Even in sections without guided instruction on the second problem, participants often floundered on the second problem and requested instruction. For these reasons, we treat the first two problems as the instructional problems and the remainder as the practice problems. We measured time to solve the whole problem, number of operations, time to perform single operators, number of operator errors, and number of transformation errors.

Results

Figure 5 shows the mean total time (time from initial presentation of the problem to successful clicking of *Next Problem* button to complete the current problem) to solve problems in the four conditions for the four chapters. The data are partitioned into performance on the first two instructional problems and performance on the remaining practice problems in each section. There are large differences in the times to solve problems for different chapters, reflecting the different number of transformations required to solve a problem in that chapter. We ignore the factor of chapter in our statistical analyses and simply use graphs like Figure 5 to show that the basic effects replicate over chapters. Therefore, our statistical analyses are 4×2 analyses of variance with the factors being the four instructional conditions and position in section (first two problems vs. later problems). In the case of total time, there are no significant effects of instructional condition, $F(3, 36) = 1.29$, $p > .25$, $MSE = 2{,}598$, or position, $F(1, 36) = 0.30$, $MSE = 554$, but there is a very strong interaction between the two, $F(3, 36) = 17.99$, $p < .0001$, $MSE = 553$. As is apparent from Figure 5, this interaction is driven by the fact that the discovery condition is worst on the initial two problems but best on the remaining problems. A contrast for this effect is highly significant, $F(1, 36) = 53.17$, $p < .0001$, whereas the residual effects in the interaction are not significant, $F(2, 36) = 0.40$. It is

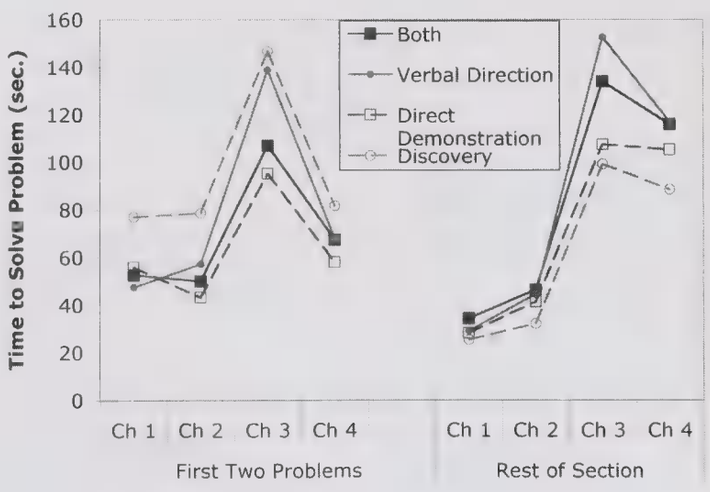


Figure 5. Time to solve problems as a function of instructional condition, chapter, and whether the problems were the first instructional problems in a section or later practice problems (Experiment 1).

not surprising that participants have difficulty on the initial couple of problems in the discovery condition. What is interesting is their superior performance on the remaining problems. These remaining problems assess what the student has learned in a section. Individual t tests confirm that the discovery condition is statistically superior to the both condition and the verbal direction conditions on the rest of the problems in the section, $t(18) = 2.78$, $p < .05$, Cohen's $d = 1.31$, and $t(18) = 3.35$, $p < .005$, $d = 1.58$, but the difference between direct demonstration and discovery does not reach significance, $t(18) = 1.40$, $p < .20$, $d = 0.66$.

The total time to solve problems can be decomposed into the number of transformations that participants perform and the time per transformation (the product of these two numbers yields the total time on a problem). These two measures are shown in Figure 6. Figure 6a shows the number of transformations and, for reference, the minimum number of transformations required for perfect performance. The main reason participants perform more than the minimum number of transformations is that they make errors in filling in the values for the transformations and have to redo them when they discover this—usually when they try to submit their answer at the end. There are main effects of condition, $F(3, 36) = 3.62$, $p < .05$, $MSE = 0.915$; and position, $F(1, 36) = 674.70$, $p < .0001$, $MSE = 0.474$; and a strong interaction between the two, $F(3, 36) = 7.17$, $p < .001$, $MSE = 0.474$. The effect of position just reflects the fact that later problems in a section tend to involve more transformations. The interaction reflects the fact that there is almost no effect of condition on the first two problems, although the conditions separate on later problems in a section. Participants did not do much more than the minimum number of transformations on the first two problems because many transformation errors are immediately flagged for these problems. On the remaining problems in the rest of the section, where transformation errors are not flagged, the discovery condition shows the fewest transformations. A contrast for this effect is highly significant, $F(1, 36) = 19.51$, $p < .0001$, whereas the residual effects in the interaction are not significant, $F(2, 36) = 1.00$. Individual t tests confirm that the discovery condition is statistically superior to all conditions on the rest of the problems in the sections: both, $t(18) = 4.05$, $p < .001$, $d = 1.91$; verbal direction, $t(18) = 3.39$, $p < .005$, $d = 1.60$; direct demonstration, $t(18) = 3.74$, $p < .005$, $d = 1.76$.

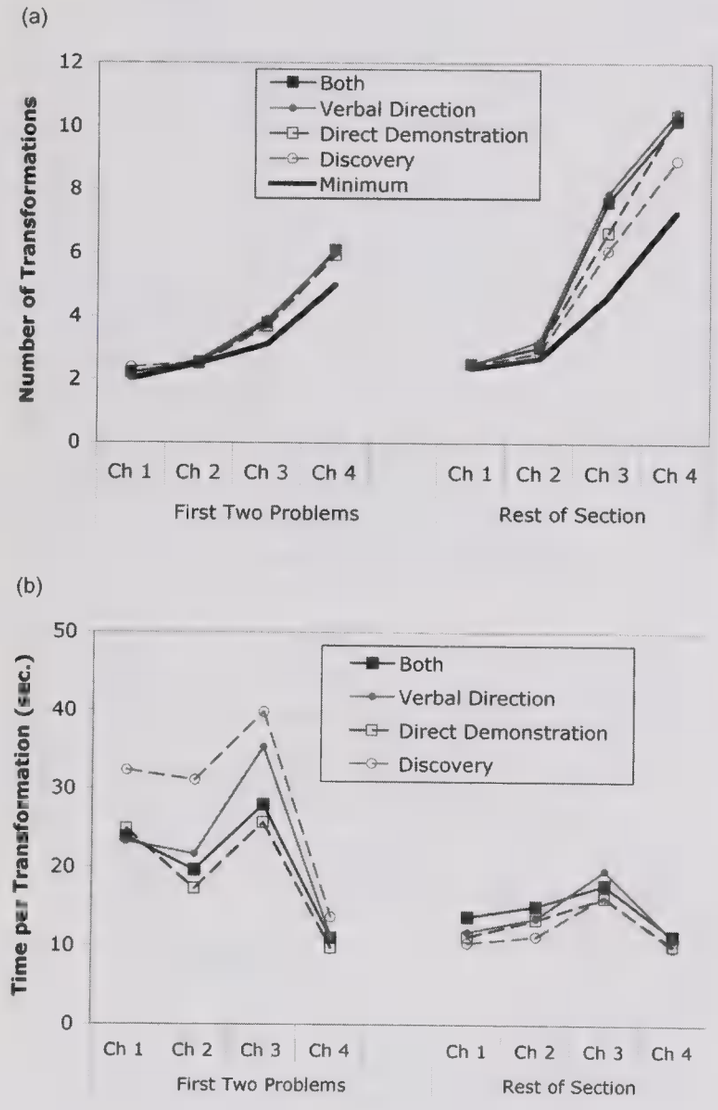


Figure 6. Mean time per transformation (a) and mean number of transformations (b) as a function of instructional condition, chapter, and whether the problems were the first instructional problems in a section or later practice problems (Experiment 1).

Figure 6b shows the time per transformation.[2] The effect of condition is not significant, $F(3, 36) = 1.69$, $p > .10$, $MSE = 170.15$, whereas the effect of position is, $F(1, 36) = 136.71$, $p < .0001$, $MSE = 55.42$. The effect of position in the section reflects a speed-up with practice. There is again a strong interaction between the two factors, $F(3, 36) = 9.62$, $p < .0001$, $MSE = 55.42$, and again this reflects the fact that the discovery condition is worst on initial problems but best for the rest of the problems in a section. Again, a contrast for this effect is highly significant, $F(1, 36) = 27.34$, $p < .0001$, whereas the residual effects are not, $F(2, 36) = 0.76$. However, this time the effect mainly comes from the slower performance of discovery students on the initial transformations, where they must find out how to perform the new transformations. This effect is particularly pronounced for the first two chapters, where most of the operations in the first problems are new. Individual t tests on the rest of the problems find no significant differences between the discovery condition and other conditions: both, $t(18) = 1.68$; verbal direction, $t(18) = 1.72$; direct demonstration, $t(18) = 0.60$; all $ps > .10$. Thus, discovery students are faster on the rest of the problems in Figure 5 because of their advantage in number of transformations (Figure 6a), not time per transformation (Figure 6b).

One can better understand the source of the difference among the conditions by considering separately the operator and transformation errors described in the introduction to this experiment. The first, the operator error, involves selecting a wrong operator for the boxes chosen (the state transitions from Figure 4b to Figure 4c and from Figure 4e to Figure 4f). This can reflect either that boxes were selected for which no operator applies or that the wrong operator was chosen for an appropriate set of boxes. These errors are flagged after the operator is chosen. The second type of error, the transformation error, involves entering the wrong values for

[2] There is a sharp drop-off in time per operation for the first two problems in chapter 4 because this chapter mainly involves putting together operations already taught to solve complex equations. Thus, with one exception, the operations in the first problems are not new.

these boxes. The tutor will accept these wrong values and transition to the next state (e.g., wrong versions of states in Figures 4d and 4g). Thus, in contrast to operator errors, transformation errors are not flagged, and students tend to go on making further operations that will eventually have to be undone. Operator errors just lengthen the duration of a transformation as the students try again for a different box–operator combination, and so they should impact the performance measure in Figure 6b. Transformation errors will increase the number of transformations in Figure 6a. Roughly stated, operator errors reflect not knowing what to do next, and transformation errors reflect not knowing how to do it. These two categories of errors are presented in Figure 7.

Figure 7a shows the mean number of operator errors per problem. The effect of condition is significant, $F(3, 36) = 6.25$, $p < .005$, $MSE = 2.67$, whereas the effect of position is not, $F(1, 36) = 1.04$, $MSE = 1.98$. There is again a strong interaction between the two, $F(3, 36) = 20.42$, $p < .0001$, $MSE = 1.98$, and this time it reflects how poorly the discovery participants were doing on the first problems where they had to discover box–operator combinations. The main effect of condition also reflects this effect on the first problems. Again, a contrast for this effect (discovery worse than the rest) on the first problems is highly significant, $F(1, 36) = 61.02$, $p < .0001$, whereas the residual effects are not, $F(2, 36) = 0.12$. Individual t tests on the rest of the problems find no significant differences between the discovery condition and other conditions: both, $t(18) = -0.16$; verbal direction, $t(18) = 0.33$; direct demonstration, $t(18) = 1.18$; all $ps > .10$.

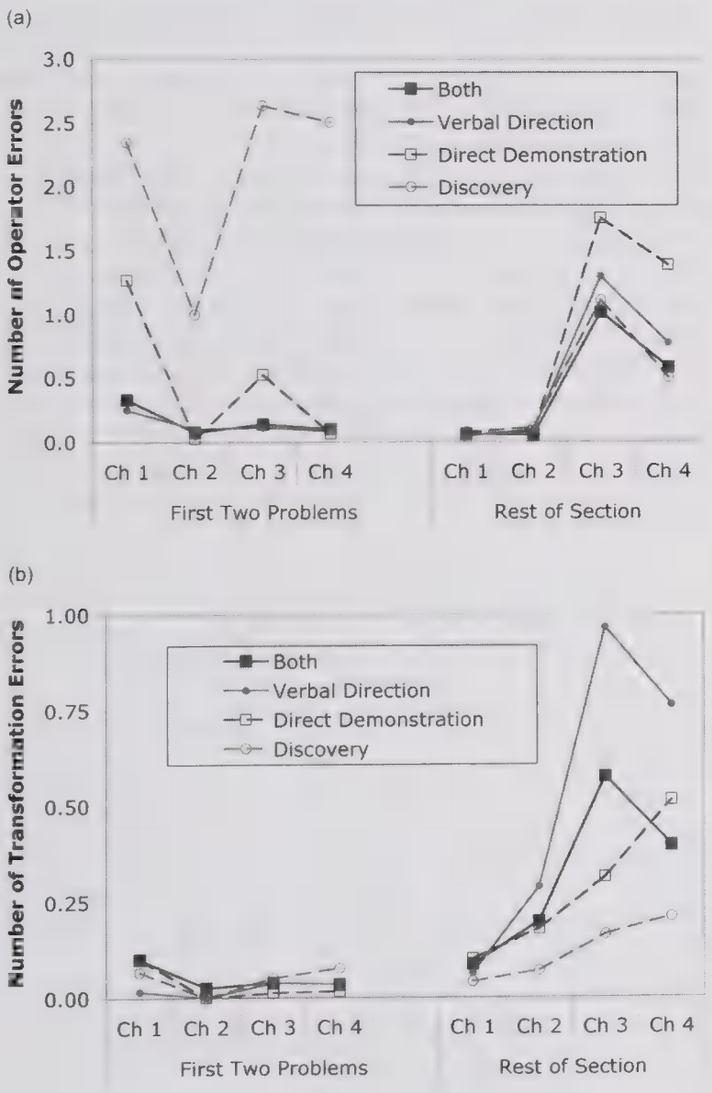


Figure 7. Mean number of operator errors (a) and transformation errors (b) as a function of instructional condition, chapter, and whether the problems were the first instructional problems in a section or later practice problems (Experiment 1).

Figure 7b shows the number of transformation errors. The effect of condition is nonsignificant, $F(3, 36) = 2.54$, $p < .10$, $MSE = 0.189$, whereas the effect of position is quite significant, $F(1, 36) = 38.16$, $p < .0001$, $MSE = 0.152$, reflecting the strong guidance provided for initial problems. The interaction of these two factors is again significant, $F(3, 36) = 4.05$, $p < .05$, $MSE = 0.152$. The interaction reflects the fact that there is almost no effect of condition in the first two problems, whereas the discovery condition is better on later problems where there is more opportunity for wrong transformations. Again, a contrast for this effect is significant, $F(1, 36) = 6.88$, $p < .05$, whereas the residual effects in the interaction are nonsignificant, $F(2, 36) = 2.64$, $p < .10$. Individual t tests confirm that the discovery condition is statistically superior to all conditions on the rest of the problems in the sections: both, $t(18) = 2.88$, $p < .01$, $d = 1.36$; verbal direction, $t(18) = 2.48$, $p < .05$, $d = 1.17$; direct demonstration, $t(18) = 2.56$, $p < .05$, $d = 1.21$.

In summary, after the first couple of learning problems the discovery condition enjoys an advantage over the other conditions on the remaining practice problems. Even if we add in the first two problems in each section, the discovery condition is at an advantage: It takes an average of 193 min to go through all 174 problems, whereas the average in the other conditions is 226 min—an advantage of over half an hour that is quite significant, $t(38) = 3.40$, $p < .005$, $d = 1.10$. Although the number of students in the conditions is not large, the effect size is very large. The advantage of the discovery condition can be traced to the fewer transformations that participants have to perform. This in turn can be traced to the fewer mistaken transformations that students make, leading to fewer repairs and less confusion.

Experiment 2

In the previous experiment, students in the discovery condition seemed to have completed their learning after the first two problems. The later problems in a section gave us evidence about what participants had learned but did not seem to be important to learning. Averaged across all sections, discovery students took 26.85 s per transformation on the first two problems, 10.89 s on the next two, and 11.87 s on the last two. Thus, there seems to be no speed-up after the first two problems in a section. The critical transformation errors were a low 2.1% per opportunity on the second two problems (it is hard to make transformation errors on the first two because of the interface) and 1.7% on the last problems. It appears that the students could have obtained the benefit of the discovery condition with far fewer problems. However, we suspected that the extra practice gave participants a familiarity with the overall system and the semantics of the diagrams that enabled them to learn so effectively in the discovery condition. To investigate this, we greatly reduced the number of problems in the second experiment, from 174 to 44. We kept the

same first two problems for each of the 12 sections but used only 20 of the remaining 152 for an average of about two extra problems per section. We tried to keep the number of extra problems approximately in proportion to the original frequency in the full set of 152. The remaining problems per section were one for section 1.1, one for section 1.2, four for section 1.7, three for section 2.6, two for section 2.7, none for section 3.1, one for section 3.2, one for section 3.4, two for section 3.5, one for section 4.1, two for section 4.2, and two for section 4.3.

The experiment was also performed to investigate a second issue about the first experiment. As Table 1 indicates, even though discovery participants did not receive any instruction about how to perform the transformations, they were given general instructions about the general purpose of the transformations—for instance, that the combine operator served to collapse boxes with two + or − operators or two × or / operators. We wanted to determine the contribution of these general instructions to learning.

There were no dramatic differences between the three instruction conditions in the first experiment. Therefore, this experiment used just one of the conditions, the direct demonstration condition, to contrast with the discovery condition. Thus, the design of the experiment crossed whether participants were given direct demonstrations or not and whether there were global instructions or not.

Method

Participants

Forty Carnegie Mellon undergraduates (27 male and 13 female; M = 23 years, *SD* = 2.1 years) took part in this study. Although they received the same performance-based feedback in terms of a financial score, the low practice in this experiment meant that students did not earn performance-based pay greater than the guaranteed minimum of $5 per half hour. Ten participants were randomly assigned to each of the four conditions produced by crossing the presence of global versus no global instructions with the factor of demonstration versus discovery. They reported relatively high algebra grades (20 *A*s, 11 *B*s, 2 *C*s, 7 missing data). These participants came from the same undergraduate pool as the first experiment, and there is no significant difference in the distribution of prior grades, $\chi^2(3, N = 80) = 2.32, p = .50$. In both cases the grade point average of the reported grades is 3.55.

Procedure

Except for fewer problems and the removal of the general instructions for half of the participants, the tutor and procedures were the same in this experiment as the previous experiment.

Results and Discussion

Qualitatively, results for the discovery conditions in this experiment differed greatly from the previous experiment. Six participants quit in the discovery condition with global instructions and four participants quit in the discovery condition without global instructions. They reached a point where they felt totally lost and did not want to continue. No participants quit in the direct demonstration conditions of this experiment, and none had quit in any conditions of the previous experiment. In addition, 3 further participants did not have enough time to complete all the problems in the discovery condition with global instructions, and 2 did not have time to complete all the problems in the discovery conditions without global directions. Thus, 50% of the discovery students quit, and another 25% went so slowly that they could not complete the experiment. In the direct demonstration condition, only one participant (without global instructions) did not complete the problems in the allotted time. The difference in number of participants completing the experiment is quite significant between the discovery and direct demonstration conditions, $\chi^2(1, N = 40) = 19.06, p < .0001$. Although there was a slightly greater tendency for greater participant loss in the discovery condition with global instructions, this was not significant, $\chi^2(1, N = 20) = 2.40, p = .12$.

Figure 8 presents the time per problem for those participants who did offer observations to a chapter (number of participants contributing is noted on the figure). Even though the poorest performing participants were eliminated on later chapters, the discovery participants were significantly worse than the direct demonstration participants at the .05 level or greater, with only one exception (the difference on the remaining problems for chapter 1). None of the differences between the two direct demonstration conditions were significant, and only one of the differences between the two discovery conditions was significant; in the rest of chapter 3, global instructions were worse than no global instructions, $t(13) = 2.23, p < .05$.

Interpreting the results for chapters 3 and 4 is problematical for another reason besides the loss of over half the participants in the discovery condition. Participants in the direct demonstration condition asked for a great many hints as they solved the rest-of-the-section problems in these chapters. In chapters 1 and 2, they averaged 0.04 hint requests per problem, whereas they averaged 3.76 for chapters 3 and 4. For comparison, instructed participants in Experiment 1 averaged 0.02 requests on the same problems for chapters 1 and 2, whereas they averaged 1.25 for chapters 3 and 4. The difference between experiments is not significant for chapters 1 and 2, $t(48) = 1.41, d = 0.40$, whereas it is highly significant for

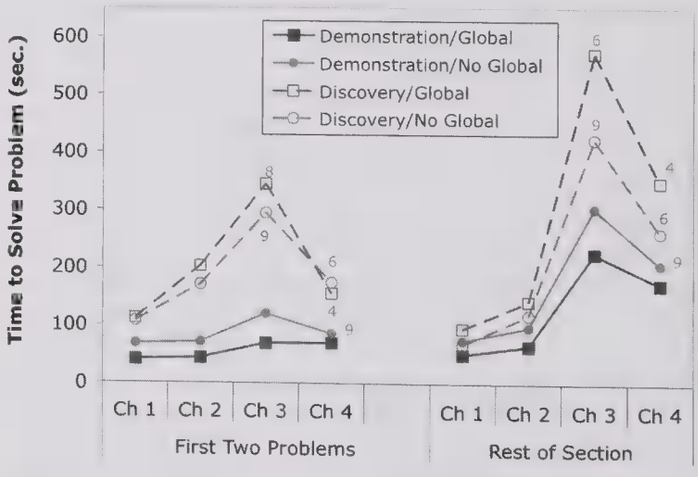


Figure 8. Time to solve problems as a function of instructional condition, chapter, and whether the problems were the first instructional problems in a section or later practice problems. The number of participants out of the original number contributing to the last two chapters is given above the data point for those chapters (Experiment 2).

chapters 3 and 4, $t(48) = 3.22$, $p < .005$, $d = 0.93$. The high rate of requests in the second experiment makes one wonder to what degree the direct demonstration participants were mastering the material in the later chapters. (Participants in the discovery condition could not ask for hints in either experiment.) Both discovery and direct demonstration participants seemed to be suffering from the lack of earlier practice when they came to these later chapters.

There were no major effects of the presence of global instructions, but there were large effects of whether the participants were in a discovery condition or were receiving directions about the individual steps of the problem. We decided to focus further analysis on this factor. Because all participants completed the first two chapters and hint requests were low for these chapters, we decided to focus on them. All the effects of the discovery condition were already in place for these two chapters. Because the effects in this experiment contrasted so sharply with the effects in the first experiment, we decided to perform a set of analyses that merged the two experiments. As the three instructional conditions of the first experiment showed few differences, we merged them into a single instruction condition and contrasted them with the discovery condition. Thus, our analysis consists of 80 participants who could be classified according to whether they were in an instruction or a discovery condition and whether they received long practice periods or short practice periods. Besides these two between-participant factors, there are the within-participant factors of chapters (1 vs. 2) and position of problem in section (first two vs. the rest). In the first two chapters, participants solved 21 problems in the short condition and 84 problems in the long condition. The first two problems were the same in the sections, and the later problems in the short condition were a subset of the later problems participants solved in the long condition. In these analyses we look only at the 21 problems that participants in both experiments solved in common.

As already noted, the students in the two experiments were drawn from the same undergraduate population, and there was no difference in their prior algebra scores. To ensure that the long and short conditions were equivalent, we looked at the first two problems for section 1.1. These problems appeared before there were any differences in practice. The mean time to solve these two problems was 57.3 s in the long instruction condition, 50.6 s in the short instruction condition, 77.9 s in the long discovery condition, and 87.9 s in the short discovery condition. The difference between instruction and discovery was highly significant, $t(76) = 3.35$, $p < .005$, $d = 0.77$, but the effect of practice length was not, $t(76) = 0.20$, $d = 0.05$, nor was the interaction between practice length and instruction, $F(1, 76) = 0.94$.

Given that the two populations are equivalent, it is significant that 50% of the participants quit the discovery condition in Experiment 2 but none did in Experiment 1. This is quite a significant difference, $\chi^2(1, N = 30) = 7.50$, $p < .01$. Nonetheless, with respect to the combined analyses that follow, it should also be noted that we are only looking at the first two chapters before there is any participant drop-out in the second experiment.

We performed the same analyses on the combined data as reported in Figures 6 and 7 for Experiment 1. Figure 9 presents the breakdown of total time to solve a problem into the number of transformations that participants perform and the time per transformation. Number of transformations (Figure 9a) shows an interaction between practice and instruction, $F(1, 76) = 5.69$, $p < .05$,

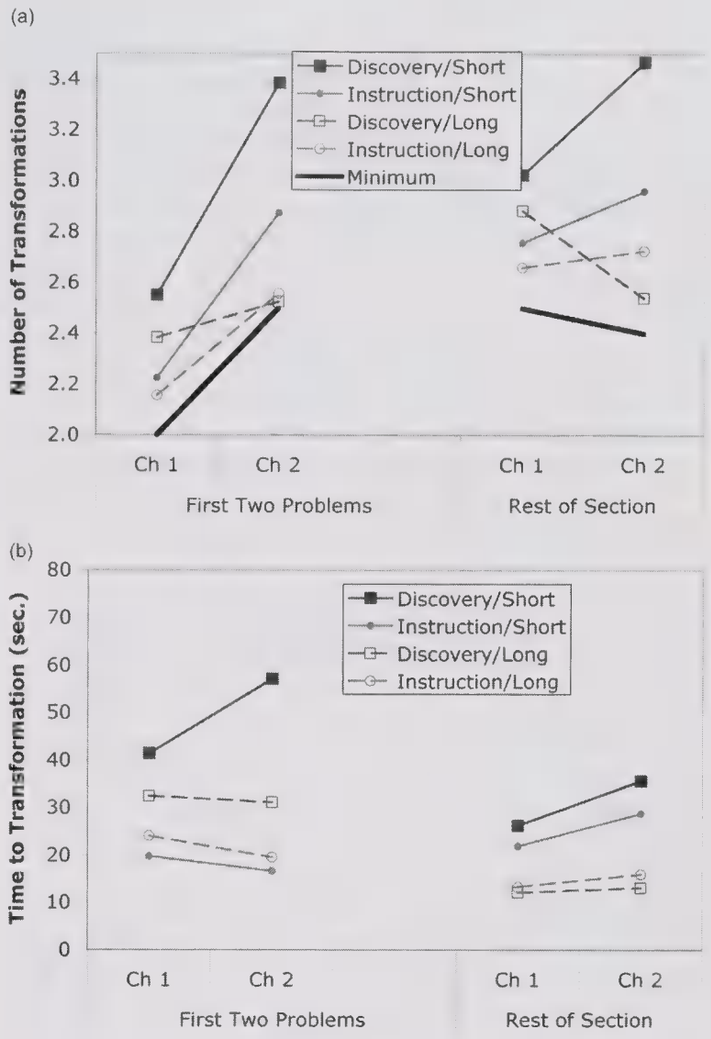


Figure 9. Mean time per transformation (a) and mean number of transformations (b) as a function of instructional condition, chapter, and whether the problems were the first instructional problems in a section or later practice problems (Experiments 1 and 2 combined).

but not an interaction between position and instruction, $F(1, 76) = 0.12$. Time per transformation (Figure 9b) shows an interaction between position and instruction, $F(1, 76) = 21.19$, $p < .0001$, but not an interaction between practice and instruction, $F(1, 76) = 0.94$. The conclusion from this figure is that the difference between the two experiments resides in the fact that discovery students in the second experiment were making a good many incorrect transformations that had to be corrected (Figure 9a).

Figure 10 shows a classification of the mean errors of the two main types. Figure 10a shows the mean number of operator errors per problem. There are strong two-way interactions between practice and instruction, $F(1, 76) = 22.65$, $p < .0001$; position and instruction, $F(1, 76) = 32.44$, $p < .0001$; and position and practice, $F(1, 76) = 15.02$; $p < .0005$. Moreover, the three-way interaction between these factors is highly significant, $F(1, 76) = 14.81$, $p < .0005$. This three-way interaction reflects the fact that participants were making many more operator errors on initial problems in the short discovery condition than any other condition. Figure 10b shows the number of transformation errors. There are two-way interactions between practice and instruction, $F(1, 76) =$

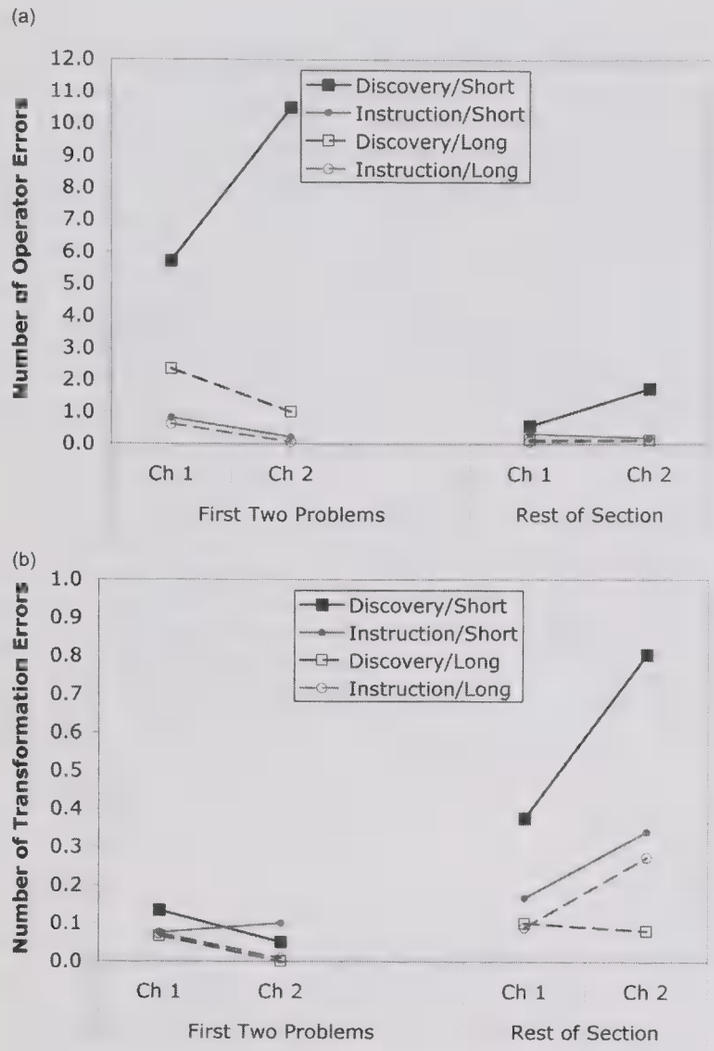


Figure 10. Mean number of operator errors (a) and transformation errors (b) as a function of instructional condition, chapter, and whether the problems were the first instructional problems in a section or later practice problems (Experiments 1 and 2 combined).

6.03, $p < .05$, and position and instruction, $F(1, 76) = 8.30$, $p < .01$. Moreover, the three-way interaction between these factors is highly significant, $F(1, 76) = 16.63$, $p < .0005$. This three-way interaction reflects the fact that participants were making many more transformation errors on later problems in the short discovery condition than in any other condition. Our characterization of this figure is that short discovery students were having much greater difficulty in identifying the correct transformations on early problems in a section (Figure 10a), and this led to a residual difficulty on later problems that shows up in transformation errors (Figure 10b). Given more practice, the students in the long condition did not have this difficulty.

Detailed Analysis of Initial Problems in Two Sections

The above analysis suggested that the difficulty of the short discovery condition began with the first problems in a section. For further insight into the initial learning in a section, Figure 11 presents a detailed analysis of behavior on the very first problems in sections 1.7 (data flow equivalent of single transformation equations) and 2.6 (combining constant terms). These two sections are distinguished by the fact that they each involve exactly two transformations; the first one is new—data-flow equivalents of rewriting $x + 3 = 8$ as $x = 8 - 3$ in section 1.7, and $3 + (2x + 7)$ as $2x + (3 + 7)$ in section 2.6—whereas the second involves the evaluation transformation (data-flow equivalents of rewriting $8 - 3$ as 5 and $3 + 7$ as 10) that they have been practicing from the beginning. (The two transformations for section 2.6 are illustrated in Figure 4.) Figure 11 displays the number of actions in excess of the minimum required taken by participants in the four conditions for each transformation. All the two-way interactions are quite significant between practice and instruction, $F(1, 76) = 13.17$, $p < .0005$; transformation and practice, $F(1, 76) = 12.01$, $p < .001$; and transformation and instruction, $F(1, 76) = 15.09$, $p < .0005$. Moreover, the three-way interaction between practice, instruction, and transformation is quite significant, $F(1, 76) = 11.54$, $p < .005$. Participants in the short discovery condition were having much greater difficulty with the first transformation than participants in any other condition and much greater difficulty with this transformation than they were having with the second transformation. Of particular note is the comparison of this group with the long discovery participants. Although the short discovery participants were somewhat worse than the long discovery participants on the second transformation, the difference is not significant, $t(28) = 1.25$, $d = 0.47$. On the other hand, the difference for the first transformation is very large and significant, $t(28) = 3.90$, $p < .001$, $d = 1.47$.

From a certain perspective, it is surprising that the short discovery participants were showing the deficit on the first transformation, which is new, and not the second transformation, which is old. We might expect that the deficit due to lack of practice would show up on the old transformation because the participants had not had as much practice with it, or that the new transformation would be equally novel to both groups and that there would have been no difference. However, the short discovery participants wandered around much more in trying to discover what they need to do to achieve the first transformation in these problems.

The important conclusions of this experiment are with respect to the discovery condition, and we do not want to make very much of the performance of students in the instruction conditions. Any

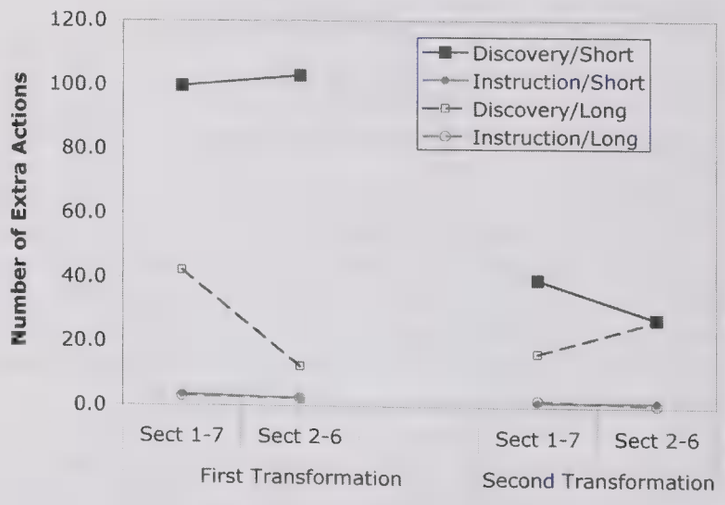


Figure 11. Mean number of actions more than the minimum on the first problem in a section as a function of instructional condition, section, and transformation (Experiments 1 and 2 combined).

difficulties instruction students had could reflect on specific properties of the instructions rather than the merit of giving instructions at all. The instruction conditions really serve as a reference point for evaluating the discovery condition. The discovery condition was superior to this reference point in the first experiment when all students had ample opportunity to practice their operators, but it was inferior to this reference point in the second experiment when this practice was removed. In contrast, practice seemed to have a much smaller beneficial effect on participants in the instruction condition (at least for the first two chapters).

All the additional practice that students received after discovering the operators for a particular section prepared them to discover operators in later sections. Students in the long discovery condition required less than half as many actions to discover the new operators as students in the short discovery condition (Figure 11). Students in the short condition had such difficulty figuring out what to do that they often seemed unable to determine how they had gotten through the problem when they finally succeeded. Therefore, they often found themselves trying to discover the operators for a section on later problems (Figure 10b). In contrast, students in the long discovery condition engaged in very little searching after the first problems of a section. The deficits begin to snowball in the short discovery condition; students were learning little from the first problems, and there were too few further problems to remedy this deficit. Thus, they were thrown into later sections without mastering the earlier sections.

Why were students so much better at guessing what to do in the long discovery condition than in the short discovery condition? We think that the practice in the long discovery condition gave students a better sense of what actions to try in a new situation, because they had developed a better understanding of the semantics of the data-flow diagrams. For instance, consider the fact (see Figure 11) that short discovery students averaged over 100 actions to find out how to achieve the first transformation for the problem in section 2.6, whereas long discovery students took fewer than 20 actions. This transformation is illustrated in the transitions between states in Figures 4a and 4d. The first correct action is to select one of the two *plus* boxes (highlighted in red after correct selection in Figure 4b). All 10 of the students in the long discovery condition selected one of these boxes as their very first action, whereas only 5 of the 20 students in the short discovery condition did. Of the remaining 15 students, 14 selected the top box with the multiplication sign (*). This then led them into a part of the problem space that was confusing until they finally backed out of it by deselecting the top box. This confusing digression led to many of the extra actions for the short discovery condition. There is no possible operation in which the top box could be involved in a useful transformation—just as there is no way to usefully transform the $2x$ in the equivalent linear expression $(2x + 7) + 3$. Students in the long condition had enough experience with these data-flow diagrams to appreciate this fact, whereas the students in the short condition were driven by superficial features like the position of the box (they had been selecting top boxes up until this point).

General Discussion

Perhaps the most important outcome of this research is the demonstration of a circumstance where discovery, with some minimal guidance, can lead to successful learning. This positive outcome depends on three factors, which were true in the long discovery condition:

1. The searches involved in making the discoveries were sufficiently constrained that it was possible for students to find solutions and remember what they had done after they discovered a successful transformation.

2. The practice enabled students to understand the semantics of these data-flow diagrams. Because the most effective way to discover operators was to use the semantics to conjecture appropriate actions, discovery students were more likely to incorporate the mathematical constraints of the domain into what they learned.

3. Because of the combinatorial nature of the problems, students had to generalize what they learned on instruction problems to novel problems. Students did better at such generalizations if they were basing actions on mathematics of the diagrams rather than superficial features like positions of boxes.

Take away any of these features (constrained search space, practice, combinatorial domain structure) and we might not have observed the superior performance of discovery students. The second experiment showed that without practice, discovery students had a very poor sense of the domain semantics. It should also be noted that these effects were obtained with a particularly able group of students. Although other results with this tutoring system have generalized from Carnegie Mellon undergraduates learning data-flow graphs to high school students learning linear equations, it remains to be shown that this result generalizes.

There are other indications in the literature that discovery can be more effective in conditions of high practice. For instance, Dean and Kuhn (2006) found in the domain of science instruction that with little practice, discovery is inferior to direct instruction (replicating Klahr & Nigam, 2004), but with extended practice it becomes equivalent or superior. Somewhat related is the expertise reversal effect (Kalyuga, Ayres, Chandler, & Sweller, 2003): More knowledgeable students require less guidance to achieve successful learning. For examples, Tuovinen and Sweller (1999) found that practice eliminated the benefit of worked examples over exploratory learning, and Kalyuga, Chandler, Tuovinen, and Sweller (2001) found that exploratory learning became superior with more practice.

With respect to general implications, this research adds to the evidence that minimally guided discovery learning can be successful if the cognitive demands are limited. One of the benefits of discovery learning is that the processes of generating a solution can lead to a characterization of the domain that will help students generalize when they face new problem situations. There is nothing magical about discovery learning in this regard and certainly not about the particular version of the discovery condition that we implemented. For instance, we expect that we would have found every bit as much advantage if participants had been instructed at every point except when they had to enter values, leaving them to discover only what to type in. This "semidiscovery" condition might have been more efficient. It has also been proposed (Aleven

& Koedinger, 2002; Roy & Chi, 2005) that the often-demonstrated advantage of self-explanation is that it encourages students to come up with correct characterization of transformations. Thus, requiring participants to generate explanations of the transformations might have been as beneficial as the discovery condition. In terms of designing instructional environments, the two critical features are that the environments do not overwhelm the cognitive resources of the student and that the discovery task encourages the student to encode the semantics that govern the combinatorial structure of the domain.

References

Aleven, V., & Koedinger, K. R. (2002). An effective meta-cognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science, 26,* 147–179.

Aleven, V., McLaren, B. M., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education, 16,* 101–128.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4,* 167–207.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research, 19,* 185–224.

Baroody, A. J., Lai, M.-L., & Mix, K. S. (2006). The development of young children's number and operation sense and its implications for early childhood education. In B. Spodek & O. Saracho (Eds.), *Handbook of research on the education of young children* (pp. 187–221). Mahwah, NJ: Erlbaum.

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review, 31,* 21–32.

Charney, D. H., Reder, L. M., & Kusbit, G. W. (1990). Goal setting and procedure selection in acquiring computer skills: A comparison of tutorials, problem-solving, and learner exploration. *Cognition and Instruction, 7,* 323–342.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18,* 293–328.

Cognitive Tutor. (2007). [Computer software]. Pittsburg, PA: Carnegie Learning.

Dean, D., & Kuhn, D. (2006). Direct instruction vs. discovery: The long view. *Science Education, 91,* 384–397.

Foerster, P. A. (1990). *Algebra I* (2nd ed.). Menlo Park, CA: Addison-Wesley.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18,* 253–292.

Fuson, K. C., Wearne, D., Hiebert, J. C., Murray, H. G., Human, P. G., Oliver, A. I., et al. (1997). Children's conceptual structures for multidigit numbers and methods of multidigit addition and subtraction. *Journal for Research in Mathematics Education, 28,* 130–162.

Heffernan, N. T., Koedinger, K. R., & Razzaq, L. (2008). Expanding the model-tracing architecture: A 3rd generation intelligent tutor for algebra symbolization. *International Journal of Artificial Intelligence in Education, 18,* 153–178.

Hiebert, J., Carpenter, T. P., Fennema, E., Fuson, K. C., Human, P., Murray, H., et al. (1996). Problem solving as a basis for reform in curriculum and instruction: The case of mathematics. *Educational Researcher, 25,* 12–21.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42,* 99–107.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). Expertise reversal effect. *Educational Psychologist, 38,* 23–31.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93,* 579–588.

Kamii, C., & Dominick, A. (1998). The harmful effects of algorithms in Grades 1–4. In L. J. Morrow & M. J. Kenney (Eds.), *The teaching and learning of algorithms in school mathematics: 1998 yearbook* (pp. 130–140). Reston, VA: National Council of Teachers of Mathematics.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41,* 75–86.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15,* 661–667.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8,* 30–43.

Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In R. K. Sawyer (Ed.), *Handbook of the learning sciences* (pp. 61–78). New York: Cambridge University Press.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59,* 14–19.

Reed, S. K., & Bolstad, C. A. (1991). Use of examples and procedures in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 753–766.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14,* 249–255.

Ritter, S., Haverty, L., Koedinger, K., Hadley, W., & Corbett, A. (2008). Integrating intelligent software tutors with the mathematics classroom. In G. Blum & K. Heid (Eds.), *Research on technology and the teaching and learning of mathematics: Vol. 2. Cases and perspectives.* Charlotte, NC: Information Age.

Rittle-Johnson, B. (2006). Promoting transfer: The effects of direct instruction and self-explanation. *Child Development, 77,* 1–15.

Roy, M., & Chi, M. T. H. (2005). Self-explanation in a multi-media context. In R. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 271–286). Cambridge, United Kingdom: Cambridge University Press.

Schmidt, H. G., Loyens, S. M. M., van Gog, T., & Paas, F. (2007). Problem-based learning is compatible with human cognitive architecture: Commentary on Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42,* 91–97.

Shulman, L. S., & Keisler, E. R. (1966). *Learning by discovery.* Chicago: Rand McNally.

Tuovinen, J. E., & Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology, 91,* 334–341.

von Glasersfeld, E. (1995). *Radical constructivism: A way of knowing and learning.* Bristol, PA: Falmer Press.

Received November 2, 2008
Revision received April 27, 2009
Accepted April 30, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 803–816

0022-0663/09/$12.00 DOI: 10.1037/a0016849

Getting a Handle on Learning Anatomy With Interactive Three-Dimensional Graphics

Andrew T. Stull, Mary Hegarty, and Richard E. Mayer
University of California, Santa Barbara

In 2 experiments, participants learned bone anatomy by using a handheld controller to rotate an on-screen 3-dimensional bone model. The on-screen bone either included orientation references, which consisted of visible lines marking its axes (orientation reference condition), or did not include such references (no–orientation reference condition). The learning task involved rotating the on-screen bone to match target orientations. Learning outcomes were assessed by asking participants to identify anatomical features from different orientations. On the learning task, the orientation reference group performed more accurately, directly, and quickly than did the control group, and high-spatial-ability individuals outperformed low-spatial-ability individuals. Assessments of anatomy learning indicated that under more challenging conditions, orientation references elevated learning by low-spatial-ability individuals to a level near that of high-spatial-ability individuals. The authors propose that orientation references assist this learning process by defining the object's main axes or providing distinguishable features.

Keywords: anatomy learning, individual differences, manual rotation, spatial cognition, virtual reality

There is a growing trend to use virtual learning resources, such as interactive three-dimensional (3-D) graphics, to augment and even replace real-world experiences in classrooms and workplaces, for example, in training of medical personnel, engineers, mechanics, skilled tradespersons, and assembly-line workers. Reducing expenses, reaching a wider audience, eliminating dangerous conditions, and coping with limited resources are often cited as justification (Bearman, 2003; Hallgren, Parkhurst, Monson, & Crewe, 2002; Reznick & MacRae, 2008). Notably, computer models and virtual experiences have been replacing cadaver and other tangible materials in the modern medical classroom (Brenton et al., 2007; Ieronutti & Chittaro, 2007; John, 2007; Nicholson, Chalk, Funnell, & Daniel, 2006; Reznick & MacRae, 2008). Occasionally stated but often assumed is the idea that virtual learning resources are as good as, or even better than, their real-world counterparts, although supporting evidence is sparse (Arnold & Farrell, 2002; Reznick & MacRae, 2008).

Despite their obvious advantages, there is some evidence that interactive 3-D graphics are difficult to use, especially for individuals with low spatial ability (Cohen & Hegarty, 2007; Keehner, Hegarty, Cohen, Khooshabeh, & Montello, 2008). For example, spatial ability is related to learning of complex and spatially demanding skills, such as surgical procedures, with virtual resources (Hegarty, Keehner, Cohen, Montello, & Lippa, 2007). Similarly, spatial ability is related to success in learning from 3-D virtual resources, such as when learning anatomy (Garg, Norman, Spero, & Maheshwari, 1999; Levinson, Weaver, Garside, McGinn, & Norman, 2007; Luursema, Verwey, Kommers, Geelkerken, & Vos, 2006).

In this set of experiments, we investigate the hypothesis that providing orientation references—visible lines marking the main axes of an object—will improve people's performance when manually rotating a virtual object (e.g., an on-screen representation of a bone) and when encoding the structure of that object. Examples of an object with and without orientation references are shown in Figure 1. We suggest that orientation references provide the learner with a *cognitive handle* that enables them to better manipulate virtual objects. We suggest that if learners are able to more efficiently manually rotate the virtual object, this frees up mental resources for developing a mental representation of the virtual object. The goals of this project were to investigate whether orientation references are helpful when manually rotating virtual objects and whether orientation references help learners develop better mental representations during anatomy learning. Because spatial ability has been shown to be related to performance in learning anatomy (Levinson et al., 2007; Rochford, 1985), we also evaluated the role of spatial ability. In particular, we were interested in whether orientation references helped low-spatial-ability learners. This research makes a theoretical contribution to our understanding of learning of complex objects and an applied contribution through improvements in the design and delivery of virtual learning resources.

To investigate learning using virtual objects, we studied a learning task in which students learned both the features of a complex, 3-D anatomical object, as illustrated in Figure 2, and the spatial

Andrew T. Stull, Mary Hegarty, and Richard E. Mayer, Department of Psychology, University of California, Santa Barbara.

This research was supported in part by National Science Foundation Grant 0313237.

We would like to acknowledge the invaluable research assistance of Bailey Bonura, Bre Gonzales, Laura Marcus, and Jana Ormsbee during this study. We also thank Jerry Tietz for his help and guidance in software programming and Paul Baker and Lewis Sadler of Visible Productions, Inc. for providing the anatomical models used in this research.

Correspondence concerning this article should be addressed to Andrew T. Stull, Department of Psychology, University of California, Santa Barbara, CA 93106. E-mail: stull@psych.ucsb.edu

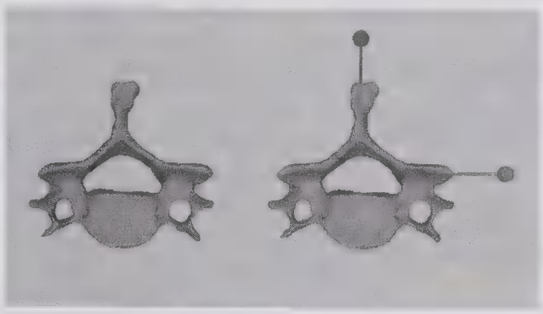


Figure 1. The model on the left shows the control condition bone. The model on the right shows the bone in the orientation reference condition, i.e., with vertical and horizontal poles. Both models are shown in a canonical orientation, the orientation used as the start position for all trials.

relations between these features. We studied this type of learning in the context of anatomy education because it represents an area where virtual learning resources are increasing in popularity. Both visual knowledge (feature identification) and spatial knowledge (spatial relationships between features) are required to form a useful mental representation of a complex anatomical object. In the learning task (manual-rotation trials), participants had to manually rotate a 3-D virtual model of a bone (a human vertebra) using a handheld interface to match a specific target orientation (as illustrated in Figure 3) and note the appearance and location of a target feature. In the course of rotating the virtual bone, we expected students to gain visual knowledge of the bone's features and spatial knowledge of the relationships between features by physically moving the virtual bone between targeted orientations. We measured knowledge after the learning trials by testing participants' ability to identify features from practiced and unpracticed orientations of the bone. We also measured how efficiently participants manipulated the virtual object during the learning trials.

The virtual lesson was intended to simulate a common teaching practice used in many anatomy classes from high school through medical school, that is, learning by exploration and manipulation of anatomical objects. This is an important area of research because technological innovations are a driving force behind many new teaching practices in medical education. Our study sought to evaluate a common learning scenario that simulated a learner

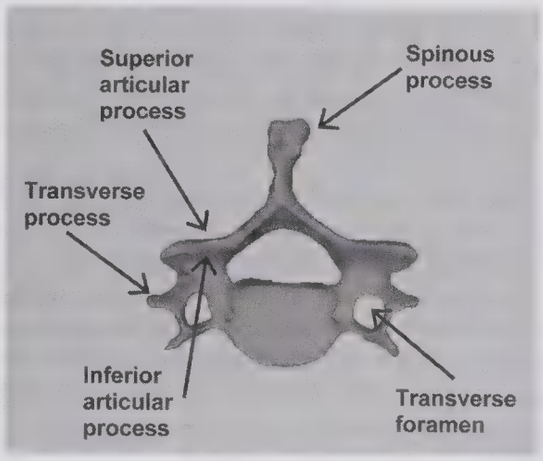


Figure 2. A model of the human cervical vertebra oriented in the starting position, indicating the five anatomical structures to be learned.

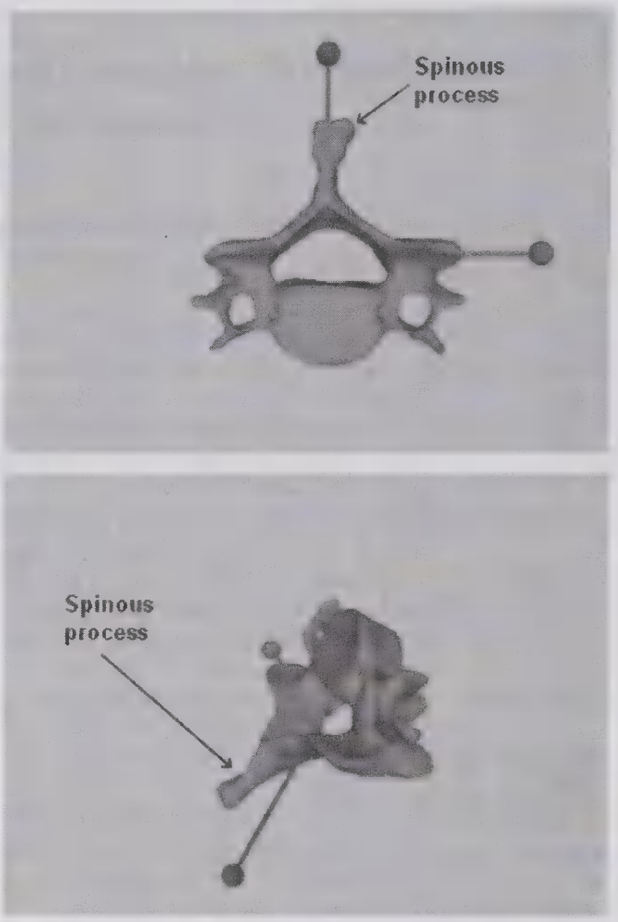


Figure 3. An example target orientation for the orientation reference condition. The top part of the page always showed the bone model in the starting orientation. The bottom part of the page always showed the bone model in the target orientation. One of the five anatomical structures (in this case, the spinous process) was indicated on each orientation of the bone in each trail.

working with a textbook illustration and a supplemental virtual object. Such a scenario creates a situation where students are challenged to develop a useful internal representation of some part of the anatomy from a static, two-dimensional (2-D) image. Traditionally, cadaver dissections and patient work help medical students develop anatomical knowledge, but these practices also carry disadvantages. Virtual materials, explored in this study, allow students to view objects from the more common noncanonical orientations that are not typical of textbook illustrations or anatomical atlases and offer a reasonable means of providing students with important learning experiences prior to their practical work with cadavers and patients.

Is Spatial Ability a Predictor of Performance of Medical Skills?

Previous research suggests that an individual's level of achievement in several medical professions, including surgery, dentistry, and nursing, is related to their spatial ability (Anastakis, Hamstra, & Matsumoto, 2000; Cuschieri, 1995; Gibbons, Baker, & Skinner, 1986; Grace, 1989; Hegarty, Keehner, Khooshabeh, & Montello, 2009; Keehner et al., 2004). Although both high- and low-spatial-ability individuals can acquire spatially demanding medical skills

with practice (Wanzel, Hamstra, Anastakis, Matsumoto, & Cusimano, 2002), spatial ability continues to predict performance after many learning sessions (Keehner, Lippa, Montello, Tendick, & Hegarty, 2006).

All medical fields rely on foundational knowledge in anatomy, which is spatially demanding in that it involves knowledge of the 3-D features and the location of anatomical structures and spatial relations between these features. Thus, it is not surprising that spatial ability has been found to be correlated with anatomy learning (Rochford, 1985). Furthermore, in medical practice, professionals are called upon to recognize and interact with anatomical structures from many different orientations, not just the canonical orientations viewed in textbooks. Virtual models of anatomy (e.g., presented via computer displays) have been advocated as a low-cost way of training anatomy that enables professionals to learn to recognize anatomical structures from different orientations. During training, students can interact with these models to view the anatomy from different orientations.

Basic research on object recognition supports the use of virtual models showing that actively controlling a virtual 3-D object during learning (as compared with passive viewing of the object) leads to more efficient recognition of the object after practice (Harman, Humphrey, & Goodale, 1999; James, Humphrey, & Goodale, 2001). However, research that used these methods to teach anatomy found that low-spatial-ability individuals had particular difficulty manipulating 3-D virtual anatomical models and had poorer learning of anatomy compared with high-spatial-ability individuals (Garg, Norman, Eva, Spero, & Sharan, 2002; Garg, Norman, & Spero, 2001; Garg, Norman, Spero, & Maheshwari, 1999). Further, Garg et al. investigated the value of learning by interacting with multiple views versus canonical views of a learned object. They concluded that there was no advantage to multiple views over canonical views and that control was not useful if canonical views were provided to the learner. However, these investigations did not consider the situation where learners need to develop spatial associations between features that are not readily visible from canonical views, a condition common in anatomy learning. We suggest that when the goal is to develop structural knowledge of a complex 3-D object, it is unlikely that one or even a few views will support the development of a coherent mental representation, without active control. We propose that when a complex task, such as medical training, requires that multiple spatially related pieces of information be derived from an object, active control is useful and, further, spatial ability is an important predictor of success.

Do Virtual Learning Resources Require Spatial Ability?

Theoretically, there are several reasons to expect that a curriculum that uses virtual learning resources will cause problems for students with lower spatial ability. Manual rotation of an object, including a virtual object, in a goal-directed task is spatially demanding because it is guided by the mental rotation of that object (Ruddle & Jones, 2001; Wexler, Kosslyn, & Berthoz, 1998; Wohlschläger, 2001; Wohlschläger & Wohlschläger, 1998). With the goal of moving an object to a desired orientation, mental rotation is used both in planning the movement and in comparing the real object with its mental representation as it is moved. Given the suggested employment of mental rotation in manipulating virtual objects, and given that low-spatial-ability individuals have difficulty with mental rotation (Hegarty & Waller, 2005), individuals with lower spatial ability are likely to be less efficient and less accurate in manually rotating virtual models of anatomy during learning.

Virtual objects burden lower spatial ability students with the need to form 3-D mental representations from 2-D representations on a computer screen. This burden is compounded by the impoverished visual and sensorimotor cues provided by virtual objects and the interface used for viewing and rotating the object (Chui et al., 2006). Thus, learning from virtual objects may be more spatially demanding than learning from real objects. This raises the question of whether it is possible to augment virtual models in ways that mitigate the challenge of using these models and developing 3-D mental representations from them.

Can Virtual Objects Be Designed to Mitigate Spatial Demands on Learning?

The identification of an object's reference frame is a common process in theories of both object recognition and mental rotation (Ballaz, Boutsen, Peyrin, Humphreys, & Marendaz, 2005; Graf, 2006; Marr, 1982; Marr & Nishihara, 1992). We propose that manual rotation of virtual objects may be particularly difficult when the object's reference frame is difficult to establish. Establishing the viewed object's reference frame may require identifying the object's main axes (Marr, 1982), identifying distinguishable features of the object (Corballis, 1988; Hayward, Zhou, Gauthier, & Harris, 2006; Mitsumatsu & Yokosawa, 2002), or both (Humphreys & Riddoch, 1984, 2006). The reference frame may be challenging to determine when the orientation of a viewed object is such that the major axes are not discernible or distinguishable features of the object are occluded. Under these circumstances, viewers might be aided by assistance with visualizing the viewed object's main axes, recognizing distinguishable features, or both.

In this article, we examine how orientation references—visible lines overlapping the object's major axes—help learners manipulate virtual models of anatomy during learning and consequently help learners develop 3-D mental representations of anatomy. Orientation references offer both a visually salient indicator of the viewed object's main axes and distinguishable features that might aid in establishing the object's reference frame. We propose that orientation references mitigate the disorientation effects that people experience when manipulating virtual objects. Lowering the effort involved in object manipulation, in turn, should allow for more effort to be invested in gaining visual and spatial knowledge of the object.

Adding orientation references, however, may provide potential disadvantages as well as advantages to the learner. One potential disadvantage may be that the orientation references act as a crutch to the learner. Orientation references are highly salient artificial devices that attract the learner's attention. Participants may attend to the orientation references to the exclusion of the relevant anatomical features that they are intended to learn. It is also possible that learners build reliance on the orientation references such that they later cannot recognize objects or form effective mental representations when orientation references are not present.

In two experiments, we investigated the effects of providing orientation references, both on manual rotation of a virtual anatomical object and on learning the structure of that object. First, we measured speed, accuracy, and directness of rotation as participants performed a manual rotation task in which they attempted to match the orientation of a virtual anatomical object to a target orientation while also noting specific anatomical features of the object (i.e., the learning phase). These three dependent measures allowed us to quantify the success (accuracy), the effort (response time), and the efficiency (directness) of manually rotating the virtual object. Second, we measured learning performance by a task in which participants had to later identify anatomical features from different orientations of the bone model (i.e., the assessment phase).

We made two predictions. First, we predicted that providing orientation references would lead to more accurate, faster, and more direct manual rotation of an object to match a target orientation. Second, we predicted that orientation references would help participants learn the anatomy. We were particularly interested in whether these predictions would hold for low-spatial-ability learners.

Experiment 1: Noncanonical Axes

In Experiment 1, we tested our predictions by asking students to match target orientations as they learned anatomical features of a bone and then take a test to identify anatomical features from different orientations. The manual rotation trials in this experiment were designed to be difficult in that they involved rotations around different noncanonical axes (i.e., axes not orthogonal to the environment or main axes of the bone) and relatively large angles of rotation (M = 130.9°, SD = 34.0°).

Method

Participants. The participants were 83 college students (M = 19.2 years, SD = 1.1) recruited from the Psychology Department Subject Pool at the University of California, Santa Barbara. Seven participants were excluded from the analysis because of equipment malfunction, experimenter error, or failure to follow directions, leaving 75 participants (30 men, 45 women) in the analysis.

Design. The study followed a 2 × 2 between-subjects design, with orientation reference (orientation reference vs. control) and spatial ability (high vs. low) as variables. Seventeen high-spatial-ability and 21 low-spatial-ability students served in the orientation reference group, and 19 high-spatial-ability and 18 low-spatial-ability students served in the control group. High- and low-spatial-ability groups were defined by a median split (Mdn = 28) of the participants' scores on the Vandenberg–Kuse Mental Rotation Test (Vandenberg & Kuse, 1978). These groups did not differ in self-reported bone anatomy knowledge. The dependent measures consisted of accuracy, response time, and directness on an object manipulation performance task, as well as accuracy on a feature identification posttest.

Materials and equipment. The materials included two versions of a manipulatable computer model of a bone, a two-page booklet describing anatomical features of the bone for the learning phase, a manual rotation battery consisting of 40 sheets of paper displaying target orientations of the bone, and 4 sheets of paper reminding students of the anatomical features of the bone during the rotation trials, a posttest consisting of 40 sheets of paper displaying bone orientations, a background knowledge questionnaire, and a mental rotation test (Vandenberg & Kuse, 1978).

The computer model was a virtual 3-D rendering of the human sixth cervical vertebra as shown in Figure 2. The model was rendered by Visible Productions, Inc., Fort Collins, Colorado from the Visual Human Project Database sponsored by the National Library of Medicine–National Institutes of Health. The model was displayed to participants with the Vizard 2.5 virtual reality program developed by WorldViz, LLC (Santa Barbara, CA), and was manipulated by participants with the InertiaCube2 three degree-of-freedom interface developed by InterSense, Inc. (Bedford, MA). The InertiaCube2 interface, sealed in a 2-in. (5.1-cm) diameter rubber ball, provided a temporal resolution of 10 ms, with an angular resolution of 0.01° root mean square deviation and was capable of measuring movement rates up to 1,200° per second.

The two versions of the bone model differed in the presence or absence of orientation references (see Figure 1). In the starting orientation, the bone model was always positioned with the dorsal spinous process at top and the right transverse process at the right side of the image. In its natural orientation within the human spine, this represents a view from below looking up toward the head. For the orientation reference condition, colored poles were added to the bone model to make the vertical and horizontal canonical axes visually salient to the participant. A blue pole extending from the top center of the bone model showed the vertical axis. A red pole extending from the center of the right side showed the horizontal axis. The two poles intersected at the bone's pivot point. Participants sat approximately 28 in. (71.1 cm) from the monitor, and the stimuli subtended a maximum visual angle of 14° for the control and 18° for the orientation reference condition but varied depending on the orientation of the virtual object. The size and length of the poles were chosen so as to make them readily viewable at diverse orientations of the bone, and the difference in visual angle was due to the addition of these poles (orientation references). The size of the bone model was identical in the two conditions.

The pretraining booklet was a two-page (8.5- by 11-in. [21.6- × 27.9-cm] sheets) description with an illustration of the anatomical features of the bone. The purpose of the booklet was to provide learners with names and general locations of structural features that they would need when rotating the virtual bone and when completing the posttest. The left-hand page included a 257-word description of the bone and five anatomical features (see Appendix). The right-hand page included a labeled illustration with the five anatomical features that were described on the left-hand page (see Figure 2). The information was printed on facing pages and could be viewed without turning the page.

Twenty target orientations illustrated on 8.5- by 11-in. sheets of paper were used for the manual rotation trials. Figure 3 illustrates the information given to participants for a typical manual rotation trial. One target orientation was represented on a single sheet with two illustrations. The top illustration always displayed the bone in the starting orientation, which was the canonical orientation, and the bottom illustration showed the bone in 1 of 20 different target orientations. Both illustrations included the name of one anatomical structure and an arrow pointing to that structure. The target orientations were determined by generating three random numbers between 0 and 359 for the separate rotation angles around the

object's three canonical axes in the order yaw (i.e., vertical axis), pitch (i.e., horizontal axis perpendicular to the viewer's line of sight), then roll (i.e., horizontal axis parallel to the viewer's line of sight). Once determined, actual target orientations were adjusted to bring both orientation reference poles into view if one or both were completely occluded by the body of the bone model. Trials were blocked, and each set of 20 target orientations was given twice in the same order, for a total of 40 trials.

Four pages containing text and diagram descriptions reminding the participant of bone anatomy were interleaved every 10 trials with the 40 target orientation pages. The purpose of these sheets was to emphasize that learning the bone anatomy was important and to provide the participants with the opportunity to verify the names of the target features. These descriptions, provided on 8.5- by 11-in. sheets, were placed at the beginning and after every 10th trial page. All descriptions were identical and illustrated the bone at the top of the page, with a 144-word text description at the bottom. The illustration was labeled with the five anatomical structures described in the two-page pretraining booklet, and the text was an abbreviated version of the description printed in the two-page booklet. The illustration for the orientation reference group included colored poles, and the illustration for the control group did not.

The 40 posttest feature identification pages, each showing one orientation of the bone, were printed on an 8.5- by 11-in. sheet of paper, as shown in Figure 4. The purpose of these sheets was to test learners' ability to recognize features from various orientations. Half of the set showed the bone in the 20 target orientations used during the rotation phase of the experiment (practiced orientations), and half of the set showed 20 bone orientations that had not been used as target orientations (unpracticed orientations). The practiced and unpracticed orientations were randomly mixed. The bone illustrations in the posttest did not include orientation references and were not labeled. Below each bone illustration was text asking the participant to circle a specific bone feature. An equal number of questions pertained to each of the five features taught in the experiment. Participants were also allowed to mark a checkbox if they could not see the requested structure or if they did not know where it was on the illustrated bone.

Figure 4. An example feature identification accuracy question.

A background knowledge questionnaire included two questions asking participants about their knowledge of bone anatomy. One question asked participants to indicate which of several anatomical structures they could identify (i.e., femur, scapula, phalanges, etc.), and the second asked participants to rate their knowledge of bone anatomy on a 5-point Likert scale. This short questionnaire was administered at the beginning of the experiment.

The Vandenberg–Kuse Mental Rotation Test is a standard measure of spatial ability with good construct validity. We selected this instrument because it most closely represents the nature of the cognitive process—depth rotation of a 3-D object—that participants were expected to perform in our experimental task.

Procedure. Participants were tested individually and were assigned alternately to the orientation reference or the control condition. The experiment was composed of three phases: (a) pretraining, (b) manual rotation trials, and (c) feature identification posttest. During the pretraining phase, participants were given written and oral instructions before completing the Vandenberg–Kuse Mental Rotation Test. Next, they were given 5 min to read a two-page anatomy training booklet describing and illustrating five anatomical structures (see Figure 2) located on the surface of the human sixth cervical vertebra. The orientation reference group received a version containing the bone illustration with colored poles (see Figure 1), and the control group received a version containing the bone illustration without the colored poles. Participants were informed that they would be tested on their knowledge of the described and illustrated information. Next, participants were given 3 min to use the InertiaCube2 interface and become comfortable with manipulating the computer model of the bone. The participants were instructed to manipulate the computer model into diverse orientations of their choosing and to identify the anatomical structures illustrated in the anatomy booklet from these diverse orientations. They were again informed that they eventually would be tested on their knowledge of the bone.

During the manual rotation phase of the experiment, participants were seated in front of the computer monitor with the 40 sheets of paper containing two sets of the 20 target orientations. Participants were instructed to use the interface to move the computer model of the bone to "quickly and accurately" match each target orientation, which was intended to minimize the use of a discovery-by-wandering strategy. Participants were asked to say the name of the target feature, which was illustrated on the target page (see Figure 3), when they were satisfied that they had matched the orientation. Naming the target feature was intended to encourage the students to learn the structural anatomy. The canonical orientation of the bone was the starting orientation for all trials. Each trial began when the participant turned to a new target orientation, at which point the experimenter pressed a key on the computer keyboard to initiate a trial. Each trial ended when the participant said the name of the anatomy structure featured on the target page, at which point the experimenter pressed a key to terminate a trial. Learners were not provided with feedback. The computer captured the time and orientation data for later analysis. Participants were given as much time as they needed to move the computer model to match the illustrated target orientation. They were informed that when they received one of the four reminder pages (spaced evenly within the

target set) they could take as much time as they wished to refresh and affirm their understanding of the bone anatomy.

During the feature identification posttest phase of the experiment, participants were given the posttest identification sheets (see Figure 4) and asked to mark with a circle or an arrow the specific location of a requested anatomical feature. Alternatively, they could check a box on the page if they could not see the requested feature or if they did not know the requested feature. Participants were allowed 30 s to complete each trial, and the next trial was presented at the end of this time period. We judged 30 s to be a sufficient time period to identify a single feature if the learners had actually learned the structural anatomy. Participants were debriefed and thanked for their participation after they completed the posttest.

Scoring. Accuracy for the manual rotation task was calculated as the average angular difference (degrees) between the target orientation and a participant's final orientation for the 40 trials (possible range = 0° to 180°).[1] Response time for the manual rotation task was calculated as time (in seconds) from a participant's first view of the target orientation to the participant's verbal statement that they had successfully matched the target orientation (averaged over 40 trials). Directness of rotation was calculated as the average integral of the object's angular distance from the target over time (degree-seconds) for 40 trials. The directness measure captures whether the participant rotated the object by the shortest path between the start orientation and the target orientation. A low value on the directness measure represents a direct and fast movement of the object, and a high value on this measure represents an indirect path, a slower movement, or both.

Feature identification accuracy was measured as the proportion correct on 37 feature identification tasks.[2] Two reviewers independently scored the identification sheets; the correlation between their scores was high, $r(76) = .91$, $p < .001$. Discrepancies in the scores between the two scorers were resolved by a third scorer.

Results and Discussion

Data analysis. Data were analyzed with separate two-way analyses of variance (ANOVAs) examining the effects of orientation references and spatial ability on each of the dependent measures: accuracy, response time, directness, and proportion of feature identification errors.[3] Table 1 lists the mean and standard deviation for each of the treatment groups on each of the four dependent measures.

Do participants with orientation references manually rotate a virtual object more accurately? The first portion of Table 1 summarizes the mean target matching accuracy (i.e., the angular deviation from the target orientation) of the four conditions in Experiment 1. The orientation reference group was significantly more accurate than the control group, $F(1, 71) = 7.62$, $MSE = 3{,}218.63$, $p = .01$, $d = 0.63$, and participants with higher spatial ability were significantly more accurate than participants with lower spatial ability, $F(1, 71) = 5.32$, $MSE = 2{,}247.88$, $p = .02$, $d = 0.46$. The interaction was not significant, $F(1, 71) = 0.05$, $MSE = 19.19$, $p = .83$. Thus, for both high- and low-spatial-ability learners, orientation references improved accuracy on the manual rotation tasks.

Do participants with orientation references manually rotate a virtual object faster? The second portion of Table 1 summarizes the mean response times of the four groups in Experiment 1. The orientation reference group responded significantly faster than the control group, $F(1, 71) = 4.31$, $MSE = 166.82$, $p = .04$, $d = 0.41$, and participants with higher spatial ability performed significantly faster than participants with lower spatial ability, $F(1, 71) = 11.05$, $MSE = 427.92$, $p < .001$, $d = 0.72$. The interaction was not significant, $F(1, 71) = 0.79$, $MSE = 30.63$, $p = .38$. Thus, for both high- and low-spatial-ability learners, orientation references improved speed on the manual rotation task.

Do participants with orientation references manually rotate a virtual object more directly? The third portion of Table 1 summarizes the mean directness of the four groups in Experiment 1. The orientation reference group was significantly more direct than the control group, $F(1, 71) = 20.02$, $MSE = 6{,}360.98$, $p < .001$, $d = 0.86$, and participants with higher spatial ability were significantly more direct than participants with lower spatial ability in moving the virtual object from the starting orientation to the target orientation, $F(1, 71) = 24.50$, $MSE = 7{,}784.09$, $p < .001$, $d = 0.79$. The interaction was not significant, $F(1, 71) = 0.28$, $MSE = 87.61$, $p = .60$. Thus, for both high- and low-spatial-ability learners, orientation references improved directness on the manual rotation task.

Does providing participants with orientation references lead to better learning of 3-D anatomy? Does learning transfer to unpracticed orientations? The fourth portion of Table 1 summarizes the mean proportion of correct feature identification questions in the posttest for the four groups in Experiment 1. A 2 × 2 × 2 mixed design ANOVA was conducted, with orientation references (orientation references vs. no orientation references) and spatial ability (high vs. low) as between-subjects variables and posttest orientation (practiced vs. unpracticed) as a within-subject variable. There was not a significant effect of orientation references, $F(1, 71) = 1.08$, $MSE = 0.02$, $p = .30$, but participants with higher spatial ability identified significantly more object features than did participants with lower spatial ability, $F(1, 71) = 13.61$, $MSE = 0.25$, $p < .001$, $d = 0.81$. These results are qualified by a significant interaction, $F(1, 71) = 5.92$, $MSE = 0.11$, $p = .02$. Contrast analyses revealed that participants with lower spatial ability in the orientation reference group correctly identified more object features, $F(1, 71) = 6.27$, $MSE = 0.06$, $p = .02$, $d = 0.76$, than did participants with lower spatial ability in the control group, but the

[1] Accuracy in matching the bone model to the target orientations was measured in *quaternions*, a base 4 hypercomplex number set used to measure the orientation of objects in three-dimensional space (Kuipers, 1999). Using quaternions, one can measure the difference between two orientations as a single value in degrees. Directness in moving the virtual bone was measured as the integral of the accuracy by time interaction for each trial. The directness measure incorporates both a spatial and a temporal component. Maximally efficient trials would consist of a minimal path of motion from the start to the target orientation, with a time course restricted only by biomechanical limits of the hand. An inefficient trial might incorporate a wandering path of motion, low accuracy in matching the target orientation, slow hand movements, or all three.

[2] Three posttest orientations (one practiced and two unpracticed) were dropped from the analysis because they could not be reliably scored.

[3] The homogeneity of variance assumption of the ANOVA was not met for the analysis of the accuracy or directness data. A rank transformation technique was applied to the data (Conover & Iman, 1981) and it was reanalyzed. Statistics for the transformed data are reported.

Table 1
Means and Standard Deviations on Manual Rotation Trials and Feature Identification by Treatment Group and Spatial Ability for Experiment 1

Measure/spatial ability	Orientation reference *M* (*SD*)	Control *M* (*SD*)
Accuracy (degree)		
High spatial ability	19.55 (8.7)	29.55 (17.8)
Low spatial ability	26.91 (15.2)	39.62 (22.1)
Response time (seconds)		
High spatial ability	12.73 (3.8)	14.44 (6.3)
Low spatial ability	16.24 (4.9)	20.52 (8.8)
Directness (degree-seconds)		
High spatial ability	857.8 (299.5)	1,204.1 (566.7)
Low spatial ability	1,175.3 (296.3)	1,911.3 (822.3)
Feature identification, total (proportion correct)		
High spatial ability	0.78 (0.11)	0.81 (0.08)
Low spatial ability	0.75 (0.06)	0.67 (0.13)
Feature identification, old (proportion correct)		
High spatial ability	0.77 (0.11)	0.79 (0.08)
Low spatial ability	0.75 (0.07)	0.66 (0.14)
Feature identification, new (proportion correct)		
High spatial ability	0.78 (0.13)	0.82 (0.08)
Low spatial ability	0.74 (0.09)	0.68 (0.14)

Note. Participants were tested on 18 practiced (old) and 19 unpracticed (new) orientations.

effect of orientation references was not significant for high-spatial-ability individuals, $F(1, 71) = 0.93$, $MSE = 0.01$, $p = .34$. Thus, orientation references improved learning of anatomy for low-spatial-ability individuals, whereas learning was good with and without orientation references for high-spatial-ability individuals.

In addition, the within-subject effect of posttest orientation (practiced vs. unpracticed) was not significant, $F(1, 71) = 1.05$, $MSE = 0.01$, $p = .31$, indicating that feature identification transferred from practiced to unpracticed orientations. The within-subject interactions with spatial ability and orientation references were also not significant. Thus, practicing with orientation references may help learners identify features from unfamiliar orientations when orientation references are not available.[4]

Further, the correlations between posttest scores and manual rotation measures for Experiment 1 showed a strong negative correlation for response time, $r(76) = -.33$, $p = .004$, and rotation directness, $r(76) = -.46$, $p < .001$, but not for rotation accuracy, $r(76) = -.15$, $p = .18$. Participants who rotated the model in less time and more directly generally did better on the posttest, but participants who more accurately matched the target orientation were not necessarily better on the posttest. These results support the idea that if people are burdened by the task of rotating the computer model, they learn less about the anatomy.

In summary, the results support our first prediction that participants rotate a virtual object more accurately, faster, and more directly when given orientation references than when not given orientation references. On average, performance of the orientation reference group was 10.8° more accurate, 2.7 s faster, and 514.8 degree-seconds more direct than the control group when manually rotating the virtual object to match a target orientation.

Further, the results show that manual rotation performance is related to spatial ability. Participants with higher spatial ability rotated a virtual object more accurately, faster, and more directly than did participants with lower spatial ability.

Finally, the results support our second prediction in that orientation references helped participants learn the anatomy of the bone. Importantly, this difference in learning attributable to orientation references was greatest for low-spatial-ability participants. Overall, participants with lower spatial ability in the orientation reference group correctly identified more features than did those in the control group.

Contrary to our concern that orientation references could act as a crutch to performance while distracting participants from attending to the intended feature information, results showed that participants were able to use the orientation references and still gain the necessary feature knowledge intended by the training. This was especially true among participants with lower spatial ability who are more likely to be burdened by added spatial content. We propose that the orientation reference effect is due to a decrease in the cognitive load (Sweller, van Merrienboer, & Paas, 1998) among participants with lower spatial ability. Without orientation references, such participants are more likely to be burdened by the task of perceiving, interpreting, and matching the orientation of a complex object while also attempting to remember the names of

[4] Posttest feature identification scores were regressed on orientation references and spatial ability. These predictors accounted for 20% of the variance in feature identification for practiced orientations, a significant effect, $F(3, 71) = 5.78$, $p = .001$. Orientation references ($b = .49$, $p = .015$), spatial ability ($b = .55$, $p < .001$), and their interaction ($b = -.44$, $p = .048$) demonstrated significant effects. For unpracticed orientations, these predictors accounted for 23% of the variance in feature identification, $F(3, 71) = 7.19$, $p < .001$. Spatial ability ($b = .55$, $p < .001$) had a significant effect in this analysis, but orientation references and their interaction did not. These results reinforce our earlier finding that lower spatial ability participants received a learning benefit from orientation references but raise the question of whether this effect transfers to unpracticed orientations.

unfamiliar features. Providing orientation references relieves the cognitive load, allowing learners to better encode object features, which contributes to the building of more complete and coherent mental representations of the attended object.

Experiment 2: Canonical Axes

Experiment 1 demonstrated that orientation references promote faster, more accurate, and more direct movement of a virtual object on the manual rotation task while allowing participants with lower spatial ability to better encode object features as indicated on the anatomy posttest. The stimuli used in Experiment 1 were generated by rotating the object by large angles ($M = 130.9°$, $SD = 34.0°$) around noncanonical axes. When the axis of rotation is not aligned with the object, the observer, or the environment, object orientations are generally more difficult to imagine and more difficult to recognize (Pani, 1993; Shiffrar & Shepard, 1991). Mental rotation is also more difficult for larger angles (Shepard & Metzler, 1971). Because of the suggested importance of mental rotation in performing manual rotation tasks (Ruddle & Jones, 2001; Wexler et al., 1998; Wohlschläger, 2001; Wohlschläger & Wohlschläger, 1998), performance in our manual rotation task should have been highly influenced by the challenging nature of the target orientations.

People may not be burdened by rotations around canonical axes and smaller angles because, evidence suggests, these conditions are generally easier to imagine (Pani, 1993). In Experiment 2, we investigated whether orientation references are helpful for these simpler rotations and whether rotations around canonical axes and smaller angles lead to equivalent mental representations of 3-D objects. Thus, in Experiment 2, we sought to replicate and extend the results of Experiment 1 by examining how effects of orientation references are moderated by the axis and angle of rotation. Overall, the manual rotation tasks were easier in Experiment 2 than in Experiment 1.

Method

Participants. The participants were 59 college students ($M = 20.0$ years, $SD = 1.7$) recruited from the Psychology Department Subject Pool at the University of California, Santa Barbara. One participant was excluded for failure to follow directions, leaving 58 participants (19 men, 39 women) in the analysis.

Design. The study followed a $2 \times 2 \times 2 \times 2$ mixed design, with orientation reference (orientation references vs. control) and spatial ability (high vs. low) as between-subjects variables and axis of rotation (canonical vs. noncanonical) and angle of rotation (small vs. large) as within-subject variables. A total of 29 students (17 high spatial ability, 12 low spatial ability) served in the orientation references group, and 29 students (17 high spatial ability, 12 low spatial ability) served in the control group. High versus low spatial ability was defined by a split of the participants' spatial ability scores on the Vandenberg–Kuse Mental Rotation Test (Vandenberg & Kuse, 1978). Participants scoring above a value of 28 were designated as having high spatial ability, and those scoring at or below 28 were considered to have low spatial ability.[5] The groups did not differ in self-reported bone anatomy knowledge. The dependent measures were the same as in Experiment 1.

Materials and equipment. The materials for Experiment 2 were identical to those for Experiment 1, with the exception of the target orientations used for the manual rotation trials. The 44 target orientations for Experiment 2 were composed of 11 rotations in 30° increments around the three canonical axes (Figure 5) and a noncanonical axis. Canonical axes comprised yaw (vertical axis), pitch (horizontal axis in the picture plane), and roll (horizontal axis parallel to the line of site). The noncanonical axis was 45° between the pitch and roll axes pointing out of the screen to the right of the observer and 45° above the horizontal plane (see Figure 5). All axes of rotation shared a common point at the origin, which was the pivot point for the virtual bone model. Small angle rotations were 30°, 60°, or 90°, and large angle rotations were 120°, 150°, and 180°.

Procedure. The procedure for Experiment 2 was identical to that of Experiment 1 except that only 15 s were allowed for each feature identification trial (because it was determined that a time of 30 s was excessive in Experiment 1).

Scoring. Accuracy, response time, directness, and feature identification accuracy were measured in the same manner as in Experiment 1. Two reviewers independently scored the identification sheets; the correlation between the scores was high, $r(58) = .94$, $p < .001$). Discrepancies between the scores of the two scorers were decided by a third rater.

Results and Discussion

Data analysis. Data were analyzed with separate $2 \times 2 \times 2 \times 2$ mixed-design ANOVAs comparing the presence and absence of orientation references, high and low spatial ability, canonical and noncanonical axes of rotation, and small and large rotation angles on each of the dependent measures: accuracy, response time, directness, and proportion correct on feature identification.[6] The means and standard deviations for the treatment groups on each of the four dependent measures are listed in Table 2.

Do participants with orientation references manually rotate a virtual object more accurately? The first portion of Table 2 summarizes the mean target matching accuracy for Experiment 2. For the between-subject effects, the orientation reference group was significantly more accurate than the control group, $F(1, 54) = 90.98$, $MSE = 0.53$, $p < .001$, $d = 2.20$, and participants with higher spatial ability were significantly more accurate than participants with lower spatial ability, $F(1, 54) = 5.88$, $MSE = 0.03$, $p = .02$, $d = 0.34$. The interaction was not significant, $F(1, 54) = 1.85$, $MSE = 0.01$, $p = .18$. As in Experiment 1, for both high- and low-spatial-ability learners, orientation references improved accuracy on the manual rotation task.

[5] The median score on the Vandenberg–Kuse Mental Rotation Test in Experiment 2 was 34. For consistency with Experiment 1, we used a score of 28 to separate higher and lower spatial ability participants in Experiment 2. The statistical results did not differ on the basis of the spatial ability grouping score.

[6] The homogeneity of variance assumption of the ANOVA was not met for the analysis of the accuracy or directness data. A reciprocal square root (1/square root of the original value) transformation was applied to the data (Field, 2005), which were then reanalyzed. Statistics for the transformed data are reported.

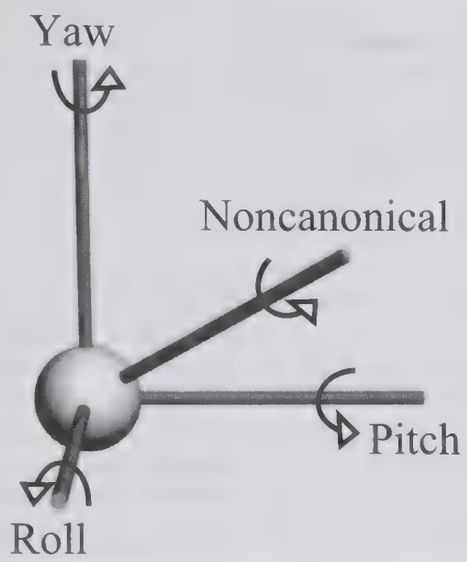


Figure 5. In Experiment 2, target orientations were developed as rotations of 30° around the three canonical axes (i.e., yaw, pitch, and roll) and a noncanonical axis.

For the within-subject effects, participants were significantly more accurate with rotations around canonical axes than with rotations around the noncanonical axis, $F(1, 54) = 47.02$, $MSE = 0.06$, $p < .001$; students were also significantly more accurate with small-angle rotations than with large-angle rotations, $F(1, 54) = 28.57$, $MSE = 0.03$, $p < .001$. Further, there was a significant interaction of axis by angle, $F(1, 54) = 16.53$, $MSE = 0.02$, $p = .02$. Contrast analyses revealed that performance on small-angle rotations was significantly more accurate than performance on large-angle rotations around the canonical axes, $F(1, 54) = 43.36$, $p < .001$, but that accuracy did not differ between small and large angles of rotations around the noncanonical axis, $F(1, 54) = 0.67$, $p = .42$.

The interaction of angle by orientation reference was also significant, $F(1, 54) = 11.25$, $MSE = 0.01$, $p = .001$. Contrast analyses revealed that orientation references eliminated the effect of angle; no difference in accuracy was evident for small- and large-angle rotations among the orientation reference group, $F(1, 54) = 1.98$, $p = .17$, whereas performance on large-angle rotations was significantly less accurate than that on small-angle rotations for the control group, $F(1, 54) = 37.83$, $p < .001$. Further, there was a significant interaction of axis by angle by orientation reference, $F(1, 54) = 5.66$, $MSE = 0.01$, $p = .02$. Contrast analyses revealed that in the orientation reference condition, accuracy was no different for large-angle rotations around canonical compared with noncanonical axes. In short, orientation references improved accuracy on difficult, but not on easy, manual rotation tasks. No other interactions were significant.

Do participants with orientation references manually rotate a virtual object faster? The second portion of Table 2 summarizes the mean response times for Experiment 2. For the between-subjects effects, the orientation reference group was not significantly faster than the control group, $F(1, 54) = 0.14$, $MSE = 9.51$, $p = .71$, and participants with higher spatial ability were not significantly faster than participants with lower spatial ability, $F(1, 54) = 2.56$, $MSE = 171.39$, $p = .12$. The interaction was not significant, $F(1, 54) = 1.53$, $MSE = 102.61$, $p = .22$. Contrary to the results of Experiment 1, orientation references did not improve speed on the manual rotation task in Experiment 2.

Within-subject effects indicated that large-angle rotations took longer than did small-angle rotations, $F(1, 54) = 27.65$, $MSE = 330.74$, $p < .001$, and that rotations around the noncanonical axis were significantly slower than rotations around the canonical axes, $F(1, 54) = 71.96$, $MSE = 754.69$, $p < .001$. Finally, the interaction of axis by angle was significant, $F(1, 54) = 6.00$, $MSE = 42.87$, $p = .02$. Contrast analyses indicate that, although performance on large-angle trials was consistently slower than performance on small-angle trials, the mean difference between large and small angles was greater for rotations around the noncanonical axis than for those around the canonical axes. No other interactions were statistically significant. These effects are consistent with known effects of axis and angle on mental rotation (Pani, 1993; Parsons, 1995; Shiffrar & Shepard, 1991).

Do participants with orientation references manually rotate a virtual object more directly? The third portion of Table 2 summarizes mean directness values for Experiment 2. For the between-subjects effects, the orientation reference group performed manual rotations more directly than did the control group, $F(1, 54) = 24.09$, $MSE = 0.01$, $p < .001$, $d = 1.20$, and participants with higher spatial ability rotated more directly than did participants with lower spatial ability, $F(1, 54) = 5.50$, $MSE = 0.001$, $p = .02$, $d = 0.65$. The interaction of orientation reference by spatial ability was not significant, $F(1, 54) = 1.43$, $MSE = 0.0003$, $p = .24$. As in Experiment 1, orientation references improved directness on manual rotation tasks for both high- and low-spatial-ability learners.

Within-subject comparisons indicated that rotations around the noncanonical axis were significantly less direct than rotations around the canonical axes, $F(1, 54) = 118.95$, $MSE = 0.004$, $p < .001$, and participants were significantly more direct in rotations around small angles than in rotations around large angles, $F(1, 54) = 31.40$, $MSE = 0.01$, $p < .001$. Further, the interaction of axis by angle was significant, $F(1, 54) = 9.13$, $MSE = 0.0002$, $p = .004$. Contrast analyses revealed that the mean difference between small and large angles was greater for rotations around the noncanonical axis (694 degree-seconds) than for rotations around the canonical axes (461 degree-seconds).

The interaction of angle by orientation reference was significant, $F(1, 54) = 9.60$, $MSE = 0.0004$, $p = .003$. Contrast analyses revealed that the mean difference between the orientation reference and control groups was greater for rotations around the large angles (478.73 degree-seconds) than for rotations around the small angles (288.61 degree-seconds). In short, orientation references improved directness to a greater extent for difficult than for easy manual rotation tasks. Finally, the interaction of angle by spatial ability was significant, $F(1, 54) = 4.83$, $MSE = 0.0002$, $p = .03$. Contrast analyses revealed that the high-spatial-ability group performed better than the low–spatial-ability group on small angles but not on large angles. No other interactions were statistically significant.

Does providing participants with orientation references lead to better learning of three-dimensional anatomy? The fourth portion of Table 2 summarizes the mean proportion correct score for feature identification in the posttest phase of Experiment 2. For the between-subjects effects, feature identification was not significantly affected by orientation references in this experiment $F(1,$

Table 2
Means and Standard Deviations on Manual Rotation Trials and Feature Identification by Treatment Group and Spatial Ability for Experiment 2

Measure, spatial ability, axis, and angle	Orientation reference *M* (*SD*)	Control *M* (*SD*)
Accuracy (degrees)		
High spatial ability		
Canonical		
Small angle	10.11 (2.58)	26.57 (13.04)
Large angle	15.03 (6.78)	44.49 (22.59)
Noncanonical		
Small angle	15.82 (4.88)	47.10 (23.13)
Large angle	14.50 (6.72)	62.01 (35.72)
Low spatial ability		
Canonical		
Small angle	14.70 (6.48)	27.28 (12.95)
Large angle	17.26 (5.68)	47.60 (16.73)
Noncanonical		
Small angle	22.15 (5.63)	48.51 (28.08)
Large angle	20.83 (12.70)	75.44 (34.27)
Response time (seconds)		
High spatial ability		
Canonical		
Small angle	8.61 (2.69)	7.86 (3.04)
Large angle	11.15 (4.79)	9.23 (2.72)
Noncanonical		
Small angle	10.16 (3.74)	10.52 (4.47)
Large angle	14.52 (7.19)	13.08 (5.34)
Low spatial ability		
Canonical		
Small angle	8.98 (4.19)	10.18 (3.43)
Large angle	10.00 (3.22)	11.46 (3.52)
Noncanonical		
Small angle	12.56 (7.57)	13.54 (5.17)
Large angle	14.48 (5.74)	17.89 (8.39)
Directness (degree-seconds)		
High spatial ability		
Canonical		
Small angle	292.27 (90.91)	414.27 (201.73)
Large angle	770.26 (345.22)	901.43 (323.23)
Noncanonical		
Small angle	439.32 (276.34)	719.56 (360.53)
Large angle	1,082.36 (659.59)	1,472.15 (721.66)
Low spatial ability		
Canonical		
Small angle	369.53 (257.81)	620.29 (314.11)
Large angle	693.80 (201.76)	1,159.28 (371.46)
Noncanonical		
Small angle	624.63 (478.54)	1,126.08 (574.43)
Large angle	1,100.55 (499.32)	2,029.05 (735.19)
Feature identification (proportion correct)		
High spatial ability	0.72 (0.07)	0.69 (0.09)
Low spatial ability	0.66 (0.17)	0.69 (0.06)

54) = 0.01, MSE = 0.00006, p = .94. Furthermore, there was no significant effect of spatial ability on feature identification in this experiment, $F(1, 54)$ = 1.11, MSE = 0.01, p = .30. Finally, the interaction of orientation references and spatial ability was not significant, $F(1, 54)$ = 1.06, MSE = 0.01, p = .31. Unlike the results of Experiment 1, orientation references did not improve performance on the anatomy posttest.[7]

The correlations between posttest scores and manual rotation measures for Experiment 2 replicated the results of Experiment 1. There were strong negative correlations of both response time, $r(58) = -.41$, p = .002, and rotation directness, $r(58) = -.36$, p = .01, with posttest performance, but again rotation accuracy, $r(58)$ = .05, p = .73, was not significantly correlated with posttest performance. Participants who rotated the model in less time and more directly generally did better on the posttest, consistent with the idea that if people are burdened by the task of rotating the computer model, they learn less anatomy.

[7] Regression of posttest feature identification scores on orientation references and spatial ability did not yield different results.

In summary, the results of Experiment 2 support our first prediction that orientation references facilitate manual rotation. When manually rotating a virtual object, participants who used orientation references were 31.08° more accurate and 383.67 degree-seconds more direct than those in the control group. In contrast to the results of Experiment 1, participants in the orientation reference group were not faster than participants in the control group on the task.

The lack of an orientation reference effect for response time may be due to the less challenging nature of the orientations used in Experiment 2. Whereas the stimuli used in Experiment 1 all involved large-angle rotations (M = 130.5°, SD = 34.0°) around multiple noncanonical axes, many of the stimuli used in Experiment 2 were small-angle rotations (30°, 60°, 90°), and three of the four axes were canonical; therefore, participants would have been expected to be faster at imagining and performing such rotations. Abundant research has shown that mental rotation around large angles or noncanonical axes are generally more challenging to perform (Pani, 1993; Parsons, 1995; Shepard & Metzler, 1971; Shiffrar & Shepard, 1991). Our results suggest that this is also the case with the manual rotation of objects under similar conditions.

The results of Experiment 2 partially replicate the results from Experiment 1 suggesting that manual rotation performance is related to spatial ability. Participants with higher spatial ability were significantly more accurate (4.77°) than participants with lower spatial ability and 203.95 degree-seconds more direct, but there was no significant difference in speed of manual rotation performance between participants with higher and lower spatial ability on the simpler rotations in this experiment.

In Experiment 2, we investigated the effects of angle (small vs. large) and axis (canonical vs. noncanonical) on rotation and learning performance. The results show that large-angle manual rotations were much more challenging than small-angle manual rotations and that rotations around noncanonical axes were much more challenging than rotations around canonical axes. For example, when rotating the object around the noncanonical axis, participants were 12.91° less accurate, 421.57 degree-seconds less direct, and 3.66 seconds slower than when rotating the object around canonical axes. In addition, participants were 10.61° less accurate, 575.37 degree-seconds less direct, and 2.43 s slower when performing large-angle rotations than when performing small-angle rotations. The results also show that the challenge of large angles and noncanonical axes was diminished for the orientation reference group, replicating the results of Experiment 1. For example, the difference in accuracy for rotations around noncanonical versus canonical axes was 21.8° in the control group but 4.0° in the orientation reference group. Similarly, the difference in accuracy for large-angle versus small-angle rotations was 20.0° in the control group but 1.2° in the orientation reference group. The benefit of orientation references is mirrored in path directness measures. The difference in directness for rotations around noncanonical versus canonical axes was 562.89 degree-seconds in the control group but 280.25 degree-seconds in the orientation reference group. These differences were all statistically significant.

Our second prediction was that providing orientation references helps people learn 3-D anatomy. The results of Experiment 2 do not replicate the results of Experiment 1. The orientation reference group did not perform significantly differently than the control group in the proportion of features correctly identified, and participants with lower spatial ability did not significantly differ from participants with higher spatial ability.

Although learning performance was generally good (average of 75% correct in Experiment 1 and 69% correct in Experiment 2), a post hoc analysis revealed a significant difference between posttest feature identification in the two experiments, $t(137) = 3.06$, $p = .003$. This difference in learning between the experiments may be due to the relationship between the orientations used for the manual rotation trials and feature identification posttest in the two experiments. In Experiment 1, the orientations used for the manual rotation phase were equally challenging to the orientations used in the feature identification posttest (i.e., large-angle rotations around unique noncanonical axes). Further, half of the 40 posttest orientations were the same as those practiced in the manual rotation trials. In Experiment 2, participants performed simpler manual rotation trials (i.e., rotations in 30° increments around three canonical axes and one noncanonical axis) and were tested with more challenging orientations in the posttest, the same orientations used in Experiment 1. The observed decrease in learning performance between Experiments 1 (75% correct) and 2 (69% correct) could have resulted from practicing with simple orientations that did not prepare the participants for testing with more challenging orientations.

An alternative interpretation is that the better learning performance of Experiment 1 is due to a practice effect because half of the posttest orientations were practiced in the rotation trials. If this interpretation were correct, then posttest accuracy should have been significantly better for practiced orientations than for unpracticed orientations, which was not the case. We interpret the results as suggesting that practicing with simple orientations did not prepare participants when challenging orientations were given on the posttest.

General Discussion

In two experiments, we examined the effects of orientation references when individuals learned anatomy by manually rotating virtual 3-D anatomical models and paying attention to labeled features of those models. The goals of the project were to investigate whether orientation references help learners manually rotate a virtual object and whether orientation references help learners develop better mental representations during anatomy learning.

Are Orientation References Helpful?

Orientation references were shown to help learners rotate virtual objects more accurately and directly in both experiments. With the more challenging trials in Experiment 1, orientation references also helped learners rotate virtual objects more quickly. Low-spatial-ability individuals learned the anatomy better with orientation references under the challenging conditions in Experiment 1, whereas learning was otherwise equivalent with and without orientation references. In Experiment 1, orientation references reduced the differences in anatomical learning between participants with higher and lower spatial ability.

For Whom Are Orientation References Helpful?

Spatial ability played a significant role in our manual rotation task. Individuals with lower spatial ability, in comparison with those of higher spatial ability, had poorer performance when rotating the virtual object. This result is consistent with previous findings that low-spatial-ability individuals have difficulty manipulat-

ing and using 3-D virtual models (Cohen & Hegarty, 2007; Garg et al., 1999; Keehner et al., 2008; Luursema et al., 2006). Although one might argue that orientation references would be beneficial primarily for performance on manual rotation tasks by low-spatial-ability individuals, in fact, orientation references were helpful to both high- and low-spatial-ability individuals. This demonstrates that all learners of all levels of spatial ability can be challenged by 3-D virtual models, and orientation references offer a method of mitigating the demands of manipulating these virtual learning resources.

Spatial ability was a contributing factor to anatomical learning, consistent with previous research (Rochford, 1985). In Experiment 1, high-spatial-ability individuals in the control condition outperformed low-spatial-ability individuals in that condition. Interestingly, this difference was reduced in the orientation reference condition, suggesting that providing these aids alleviated difficulties faced by low-spatial-ability individuals in learning anatomy.

The poorer performance of lower spatial ability participants in the control conditions of these experiments may be due to difficulty recognizing the orientation of an object when its main axis is foreshortened or when distinguishable features are occluded. This effect may also be due to poor ability to mentally rotate the perceived object for comparison with a representation of that object from a different orientation. Finally, poor performance may be due to lower ability to mentally compare features in representations of the same object from different orientations. In our experiments, the comparison in manual rotation was one between different views of the object that were presented externally, whereas the comparison in the posttest feature recognition task occurred between a view of the object presented externally and the participant's internal mental representation. We propose that by providing orientation references, which clearly mark the orientation of the object, we reduced participants' cognitive load during the learning (manual rotation) phase of the experiment and consequently enabled them to construct more coherent mental representations, which they could then use to recognize features during the posttest trials.

When Are Orientation References Helpful?

A comparison of the results of Experiments 1 and 2 reveals that orientation references are most helpful under the most challenging conditions for both manual rotation and anatomical learning. First, orientation references had larger effects on both manual rotation and learning in Experiment 1 (which used large angles and noncanonical axes) than in Experiment 2. Second, in Experiment 2, orientation references helped more for rotations around noncanonical axes than for rotations around canonical axes. Third, in Experiment 2, orientation references helped more for large angles than for small angles.

Orientation references are therefore most helpful under conditions that are typical of medical practice. Medical professionals such as surgeons, radiologists, and nurses are often called upon to recognize anatomical structures from noncanonical orientations. Our research suggests that providing orientation references during learning will allow professionals to construct coherent mental representations that enable them to recognize features of anatomical structures from diverse perspectives. However, more work is needed to understand how this technique is effective over a long-term period.

It is interesting that manual rotation of virtual objects is affected by axis and angle of rotation, which are also performance challenges associated with mental rotation (Pani, 1993; Parsons, 1995; Shepard & Metzler, 1971; Shiffrar & Shepard, 1991). This supports the view that mental rotation is a component of manual rotation (Ruddle & Jones, 2001; Wexler et al., 1998; Wohlschläger, 2001; Wohlschläger & Wohlschläger, 1998). The fact that manual rotations around large angles and noncanonical axes were less accurate is somewhat surprising given that participants in our study had more than enough time and opportunity to perform and validate their actions.

How Are Orientation References Helpful?

Given the positive effects of orientation references in our experiments, it is important to consider the mechanisms by which they confer benefit. Our hypotheses were based on the importance of establishing an object's reference frame during recognition (Ballaz et al., 2005; Corballis, 1988; Graf, 2006; Hayward et al., 2006; Humphreys & Riddoch, 1984, 2006; Marr, 1982; Marr & Nishihara, 1992; Mitsumatsu & Yokosawa, 2002). This aspect of object recognition is relevant to both manual rotation of a virtual anatomical object and recognizing features of that object from different orientations. Our orientation references may aid participants by either (a) defining the main axes of the object or (b) offering visually salient and easily distinguishable features. The results of our study do not distinguish between these hypotheses but are consistent with both. Future research should evaluate whether visually salient and easily distinguishable features would be equally effective as orientation references if they did not define the main axes of the object.

Implications

Our research suggests that virtual learning resources may increase rather than diminish the burden imposed on low-spatial-ability learners in spatially demanding professions. Poorly designed virtual resources can impose an unnecessary yet preventable disadvantage for individuals who, if given adequate aids, may develop into successful practitioners. The orientation reference technique we explored in this research is an example of one way to minimize problems of low-spatial-ability learners when using virtual resources. With the spatial burden lightened by cognitive handles, such as the orientation references explored here, individuals may be better able to develop the skills and knowledge that they need to be successful in spatially demanding careers.

References

Anastakis, D. J., Hamstra, S. J., & Matsumoto, E. D. (2000). Visual–spatial abilities in surgical training. *American Journal of Surgery, 179,* 469–471.

Arnold, P., & Farrell, M. J. (2002). Can virtual reality be used to measure and train surgical skills? *Ergonomics, 45,* 362–379.

Ballaz, C., Boutsen, L., Peyrin, C., Humphreys, G. W., & Marendaz, C. (2005). Visual search for object orientations can be modulated by canonical orientation. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 20–39.

Bearman, M. (2003). Is virtual the same as real? Medical students' experiences of a virtual patient. *Academic Medicine, 78,* 528–545.

Brenton, H., Hernandez, J., Bello, F., Strutton, P., Purkayastha, S., Firth, T., & Darzi, A. (2007). Using multimedia and Web3D to enhance anatomy teaching. *Computers and Education, 49,* 32–53.

Chui, C., Ong, J. S. K., Lian, Z., Want, Z., Teo, J., Zhang, J., et al. (2006). Haptics in computer-mediated simulation: Training in vertebroplasty surgery. *Simulation and Gaming, 37,* 438–451.

Cohen, C. A., & Hegarty, M. (2007). Individual differences in use of

external visualizations to perform an internal visualization task. *Applied Cognitive Psychology, 21,* 701–711.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician, 35,* 124–129.

Corballis, M. C. (1988). Recognition of disoriented shapes. *Psychological Review, 95,* 115–123.

Cuschieri, A. (1995). Visual displays and visual perception in minimal access surgery. *Seminars in Laparoscopic Surgery, 2,* 209–214.

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.

Garg, A. X., Norman, G. R., Eva, K. W., Spero, L., & Sharan, S. (2002). Is there any real virtue of virtual reality? The minor role of multiple orientations in learning anatomy from computers. *Academic Medicine, 77,* S97–S99.

Garg, A. X., Norman, G. R., & Spero, L. (2001). How medical students learn spatial anatomy. *The Lancet, 357,* 363–364.

Garg, A. X., Norman, G. R., Spero, L., & Maheshwari, P. (1999). Do virtual computer models hinder anatomy learning. *Academic Medicine, 74,* S87–S89.

Gibbons, R. D., Baker, R. J., & Skinner, D. B. (1986). Field articulation testing: A predictor of technical skills in surgical residents. *Journal of Surgical Residence, 41,* 53–57.

Grace, D. M. (1989). Aptitude testing in surgery. *Canadian Journal of Surgery, 32,* 396–397.

Graf, M. (2006). Coordinate transformations in object recognition. *Psychological Bulletin, 132,* 920–945.

Hallgren, R. C., Parkhurst, P. E., Monson, C. L., & Crewe, N. M. (2002). An interactive, Web-based tool for learning anatomic landmarks. *Academic Medicine, 77,* 263–265.

Harman, K. L., Humphrey, G. K., & Goodale, M. A. (1999). Active manual control of object views facilitates visual recognition. *Current Biology, 9,* 1315–1318.

Hayward, W. G., Zhou, G., Gauthier, I., & Harris, I. M. (2006). Dissociating viewpoint costs in mental rotation and object recognition. *Psychonomic Bulletin and Review, 13,* 820–825.

Hegarty, M., Keehner, M., Cohen, C., Montello, D. R., & Lippa, Y. (2007). The role of spatial cognition in medicine: Applications for selecting and training professionals. In G. L. Allen (Ed.), *Applied spatial cognition: From research to cognitive technology* (pp. 285–315). Mahwah, NJ: Erlbaum.

Hegarty, M., Keehner, M., Khooshabeh, P., & Montello, D. R. (2009). How spatial ability enhances, and is enhanced by, dental education. *Learning and Individual Differences, 19,* 61–70.

Hegarty, M., & Waller, D. A. (2005). Individual differences in spatial abilities. In P. Shah & A. Miyake (Eds.), *Handbook of visuospatial thinking.* Cambridge, United Kingdom: Cambridge University Press.

Humphreys, G. W., & Riddoch, M. J. (1984). Routes to object constancy: Implications from neurological impairments of object constancy. *Quarterly Journal of Experimental Psychology, 36*A, 385–415.

Humphreys, G. W., & Riddoch, M. J. (2006). Features, objects, action: The cognitive neuropsychology of visual object processing, 1984–2004. *Cognitive Neuropsychology, 23,* 156–183.

Ieronutti, L., & Chittaro, L. (2007). Employing virtual humans for education and training in ×3D/VRML worlds. *Computers and Education, 49,* 93–109.

James, K. H., Humphrey, G. K., & Goodale, M. A. (2001). Manipulating and recognizing virtual objects: Where the action is. *Canadian Journal of Experimental Psychology, 55,* 111–120.

John, N. W. (2007). The impact of Web3D technologies on medical education and training. *Computers and Education, 49,* 19–31.

Keehner, M. M., Hegarty, M., Cohen, C., Khooshabeh, P., & Montello, D. R. (2008). Spatial reasoning with external visualizations: What matters is what you see, not whether you interact. *Cognitive Science, 32,* 1099–1132.

Keehner, M., Lippa, Y., Montello, D. R., Tendick, F., & Hegarty, M. (2006). Learning a spatial skill for surgery: How the contributions of abilities change with practice. *Applied Cognitive Psychology, 20,* 487–503.

Keehner, M. M., Tendick, F., Meng, M. V., Anwar, H. P., Hegarty, M., Stoller, M. L., & Duh, Q. (2004). Spatial ability, experience, and skill in laparoscopic surgery. *The American Journal of Surgery, 188,* 71–75.

Kuipers, J. B. (1999). *Quaternions and Rotation Sequences: A primer with applications to orbits, aerospace, and virtual reality.* Princeton, NJ: Princeton University Press.

Levinson, A. J., Weaver, B., Garside, S., McGinn, H., & Norman, G. R. (2007). Virtual reality and brain anatomy: A randomised trial of e-learning instructional designs. *Medical Education, 41,* 495–501.

Luursema, J., Verwey, W. B., Kommers, P. A. M., Geelkerken, R. H., & Vos, H. J. (2006). Optimizing conditions for computer-assisted anatomical learning. *Interacting with Computers, 18,* 1123–1138.

Marr, D. (1982). The philosophy and the approach. In S. Yantis (Ed.), *Visual perception: Essential readings* (pp. 104–123). New York: Psychology Press.

Marr, D., & Nishihara, H. K. (1992). Visual information processing: Artificial intelligence and the sensorium of sight. In S. M. Kosslyn & R. A. Andersen (Eds.), *Frontiers in cognitive neuroscience* (pp. 165–186). Cambridge, MA: MIT Press.

Mitsumatsu, H., & Yokosawa, K. (2002). How do the internal details of the object contribute to recognition? *Perception, 31,* 1289–1298.

Nicholson, D. T., Chalk, C., Funnell, W. R. J., & Daniel, S. J. (2006). Can virtual reality improve anatomy education? A randomized controlled study of a computer-generated three-dimensional anatomical ear model. *Medical Education, 40,* 1081–1087.

Pani, J. R. (1993). Limits on the comprehension of rotational motion: Mental imagery of rotations with oblique components. *Perception, 22,* 785–808.

Parsons, L. M. (1995). Inability to reason about an object's orientation using an axis and angle of rotation. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 1259–1277.

Reznick, R. K., & MacRae, H. (2008). Teaching surgical skills—Changes in the wind. *The New England Journal of Medicine, 355,* 2664–2669.

Rochford, K. (1985). Spatial learning disabilities and underachievement among university anatomy students. *Medical Education, 19,* 13–26.

Ruddle, R. A., & Jones, D. M. (2001). Manual and virtual rotations of a three-dimensional object. *Journal of Experimental Psychology: Applied, 7,* 286–296.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171,* 701–703.

Shiffrar, M. M., & Shepard, R. N. (1991). Comparison of cube rotations around axes inclined relative to the environment or to the cube. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 44–54.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10,* 251–296.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills, 47,* 599–601.

Wanzel, K. R., Hamstra, S. J., Anastakis, D. J., Matsumoto, E. D., & Cusimano, M. D. (2002). Effect of visual–spatial ability on learning of spatially-complex surgical skills. *Lancet, 359,* 230–232.

Wexler, M., Kosslyn, S. M., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition, 68,* 77–94.

Wohlschläger, A. (2001). Mental object rotation and the planning of hand movement. *Perception & Psychophysics, 63,* 709–718.

Wohlschläger, A., & Wohlschläger, A. (1998). Mental and manual rotation. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 397–412.

(Appendix follows)

Appendix

Anatomical Features of the Human Sixth Cervical Vertebrae

The individual bones of your spinal column are the **vertebrae** (plural for *vertebra*). Each vertebra forms a bony ring, which creates a hollow tube when stacked one on top of another. Your vertebrae surround and protect your spinal cord and also provide a framework where muscles attach and where other bones join. Your muscles allow mobility and your joints allow flexibility.

Vertebrae have many parts that are divided into general structures called processes and foramina (plural for *foramen*). A **process** is a location on a bone where a muscle attaches or another bone meets to form a joint. A **foramen** is an opening or passage for nerves or blood vessels.

- The **spinous process** is the long bony projection on each vertebra. You can feel the spinous process on each bone if you run your hand down your back. The spinous process is just one location where your muscles attach to the vertebrae.
- The two **transverse processes** are on either side of your vertebrae and are locations where additional muscles attach.
- The **superior articular process** and **inferior articular process** form joints between adjacent vertebrae. The superior (upper) articular process of one vertebra meets the inferior (lower) articular process of an adjacent vertebra. These processes overlap each other to form two flexible joints, one on either side of the vertebrae.
- The **transverse foramen** is the opening on either side of your vertebrae. These two openings are where nerves bundles enter and exit your spinal cord to reach the rest of your body.

Received November 20, 2008
Revision received April 27, 2009
Accepted June 23, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 817–835

0022-0663/09/$12.00 DOI: 10.1037/a0016127

Spatial Ability for STEM Domains: Aligning Over 50 Years of Cumulative Psychological Knowledge Solidifies Its Importance

Jonathan Wai, David Lubinski, and Camilla P. Benbow
Vanderbilt University

The importance of spatial ability in educational pursuits and the world of work was examined, with particular attention devoted to STEM (science, technology, engineering, and mathematics) domains. Participants were drawn from a stratified random sample of U.S. high schools (Grades 9–12, N = 400,000) and were tracked for 11+ years; their longitudinal findings were aligned with pre-1957 findings and with contemporary data from the Graduate Record Examination and the Study of Mathematically Precocious Youth. For decades, spatial ability assessed during adolescence has surfaced as a salient psychological attribute among those adolescents who subsequently go on to achieve advanced educational credentials and occupations in STEM. Results solidify the generalization that spatial ability plays a critical role in developing expertise in STEM and suggest, among other things, that including spatial ability in modern talent searches would identify many adolescents with potential for STEM who are currently being missed.

Keywords: spatial ability, talent searches, longitudinal study, STEM, constructive replication

Over 50 years ago, Super and Bachrach (1957) published *Scientific Careers*, a report of a National Science Foundation (NSF) advisory panel. Appearing the year *Sputnik* was launched, this document characterized the personal attributes of scientists and engineers for the purposes of better identifying human capital and, ultimately, uncovering ways to nurture scientific and technical potential. It also was the year of two landmark publications in the *American Psychologist*: Cronbach's (1957) APA Presidential Address, on "The Two Disciplines of Scientific Psychology," wherein the importance of tailoring educational interventions and opportunities to individual differences among students was emphasized, and Paterson's (1957) Bingham Lecture, "The Conservation of Human Talent," which reinforced this idea.

Emphasized throughout Super and Bachrach (1957) was the critical role of spatial ability, a construct aptly defined by Lohman (1994a, p. 1000) as "the ability to generate, retain, retrieve, and transform well-structured visual images." Spatial ability was characterized as an individual differences attribute with particular relevance for learning the advanced scientific–technical material needed for developing outstanding STEM (science, technology, engineering, and mathematics) contributors, those individuals capable of moving engineering and physical science disciplines forward. However, in their review Super and Bachrach stressed that attributes beyond spatial ability—mathematical ability in particular, as well as interests and nonintellectual determinants such as persistence—should be studied also. They further voiced that "longitudinal studies beginning at a relatively early age and extending over a period of some 10 to 15 years seemed called for" (Super & Bachrach, 1957, p. 87). This study sequences two such longitudinal studies: one from 1960 to 1974 and a second that began in 1971 and is still ongoing.

Contemporary Neglect of Utilizing Psychological Knowledge About Spatial Ability

Part of the motivation for this article is that currently, over 50 years after Super and Bachrach's (1957) report, relatively little implementation of spatial ability is found for selection, curriculum, and instruction in educational settings—even in STEM domains, where it appears to be highly relevant. This neglect is especially surprising as we live in a globally competitive world (Friedman, 2005), and the need to identify and nurture scientific and technical talent has never been greater (American Competitiveness Initiative, 2006; National Academy of Sciences, 2005). Indeed, with plenty of evidence for the educational–occupational significance of spatial ability accumulated (Gohm, Humphreys, & Yao, 1998; Humphreys, Lubinski, & Yao, 1993; Lohman, 1988, 1994a, 1994b; Smith, 1964), Richard E. Snow (1999) expressed perplexity about the neglect of spatial ability in applied educational circles:

> There is good evidence that [spatial ability] relates to specialized achievements in fields such as architecture, dentistry, engineering, and medicine . . . . Given this plus the longstanding anecdotal evidence on the role of visualization in scientific discovery, . . . it is incredible that

Jonathan Wai, David Lubinski, and Camilla P. Benbow, Department of Psychology and Human Development, Vanderbilt University.

Support for this article was provided by a research and training grant from the Templeton Foundation, the Society of Multivariate Experimental Psychology, and National Institute of Child Health and Development Grant P30 HD 15051 to the Vanderbilt Kennedy Center for Research on Human Development. Earlier versions of this article benefited from comments from Kimberley Ferriman, Linda S. Gottfredson, Gregory Park, Stijn Smeets, and Maya Wai.

Correspondence concerning this article should be addressed to Jonathan Wai, David Lubinski, or Camilla P. Benbow, Department of Psychology and Human Development, Vanderbilt University, 0552 GPC, 230 Appleton Place, Nashville, TN 37203. E-mail: jonathan.wai@vanderbilt.edu, david.lubinski@vanderbilt.edu, or camilla.benbow@vanderbilt.edu

there has been so little programmatic research on admissions testing in this domain. (p. 136)

Since Snow's (1999) observation, at least two promising studies have appeared that further underscore the importance of assessing spatial ability among intellectually talented youths initially identified by mathematical and verbal measures. These studies also suggest a venue wherein assessing spatial ability could have an immediate impact, because both were based on talent search participants (Benbow & Stanley, 1996; Colangelo, Assouline, & Gross, 2004; Stanley, 2000). (Talent search participants are young adolescents who take college entrance exams 4 years earlier than is typical in order to qualify for special educational programs for talented youths.) Talent searches could relatively easily add spatial ability measures to their selection criteria and thereby cast a wider net for identifying intellectually able youths for educational experiences in architecture, engineering, robotics, and the physical sciences. However, the assessment of spatial ability may benefit more students than just talented youths. Basic science indicates that students throughout the ability range could profit from spatial ability assessments and the provision of educational opportunities aimed at developing spatial ability (Humphreys et al., 1993; Humphreys & Lubinski, 1996; Lohman, 2005; Smith, 1964).

The two studies on spatial ability discussed above that appeared after Snow (1999) were based on independent cohorts of participants in the Study of Mathematically Precocious Youth (SMPY; Lubinski & Benbow, 2006). SMPY is a longitudinal study currently in its fourth decade and consisting of five cohorts identified at different time points. It is designed to uncover the best methods for identifying and nurturing talent for STEM as recommended by Super and Bachrach (1957). Shea, Lubinski, and Benbow (2001) tracked 563 talent search participants identified with the Scholastic Assessment Test (SAT) by age 13 as intellectually talented (top 0.5% for their age-group); at the time of their identification in the late 1970s, they were assessed on spatial ability also. Over a 20-year interval, biographical, educational, and occupational criteria were collected 5, 10, and 20 years after initial identification. Relative to criterion groupings in the humanities and other disciplines, the young adolescents who subsequently found math–science to be their favorite high school course, earned undergraduate and graduate degrees in STEM, and ultimately ended up in a STEM career 20 years later, typically displayed higher levels of spatial ability at age 13. Moreover, the discriminant function analyses conducted at all three time points revealed that spatial ability added *incremental validity* (accounted for a statistically significant amount of additional variance) beyond SAT-Mathematical (measuring mathematical reasoning ability) and SAT-Verbal (measuring verbal reasoning ability) in predicting these math–science criteria.

Subsequently, Webb, Lubinski, and Benbow (2007), using a less select talent search sample of 1,060 adolescents identified in the mid-1990s (top 3% in ability), provided evidence that spatial ability possesses incremental validity over both SAT scales and comprehensive educational–occupational preference questionnaires over a 5-year interval for predicting favorite high school course, leisure activities relevant to STEM, college major, and intended occupation (i.e., the predictive period spanned from initial identification at age 13 to after high school). Overall, spatial ability accounted for an additional 3% of the variance in predicting these criteria beyond both SAT measures and two comprehensive educational–vocational preference questionnaires. Again, relative to young adolescents whose outcomes fell in the humanities and in other disciplines, participants with STEM outcomes displayed higher levels of spatial ability at age 13.

As suggestive as these findings are, however, D. F. Lohman (personal communication, May 2007) noted one limitation: Shea et al. (2001) and Webb et al. (2007) were not based on random samples of the general population or even random samples of high-ability students. All participants in both studies were talent search participants, students identified as highly able who often were motivated to attend academically challenging programs for talented youths. Would spatial ability play a similar role among students not identified in this fashion? One purpose in our study is to provide an answer to this question. In addition, to solidify the length of time that spatial ability has been known to play a consistent role in the development of STEM expertise, we decided to try to bridge the gap between the studies reviewed in Super and Bachrach's (1957) NSF report and our contemporary findings from talent search participants (Shea et al., 2001; Webb et al., 2007). For this purpose, we explored Project TALENT, a massive longitudinal study launched just following Super and Bachrach's report (in 1960) and culminating with an 11-year follow-up in the early 1970s (Wise, McLaughlin, & Steel, 1979), when the first SMPY participants were identified. Figure 1 illustrates the bridge we are aiming to build.

Project TALENT consists of four cohorts totaling 400,000 participants. They were identified as high school students (Grades 9 through 12, approximately 100,000 per grade) shortly after Sputnik was launched. Subsequently, they were followed up 11 years after their high school graduation in the early 1970s, when modern talent searches for intellectually precocious youths were just being launched (Keating & Stanley, 1972; Stanley, 1996). Therefore, if the findings uncovered by this study of Project TALENT participants correspond with those of studies conducted prior to and reviewed in Super and Bachrach's (1957) NSF report, and, in addition, if they mirror modern findings based on talent search participants identified throughout the 1970s and 1990s and fol-

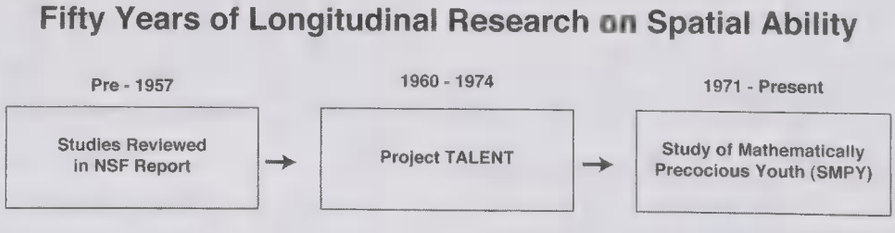


Figure 1. Over 50 years of cumulative empirical research on the educational and occupational significance of spatial ability for STEM.

lowed up in current times, the collective findings would establish a solid foundation for educational practice. This foundation would be derived from two distinctive longitudinal studies of the type Super and Bachrach (1957) called for (and which were launched over successive time frames): Project TALENT (1960 to early 1970s) and SMPY (early 1970s to present times).

Studies that have been conducted periodically for over 50 years with different populations and that consistently reveal similar patterns are rare in educational psychology. Furthermore, there is a methodological rationale for the importance of sequencing such studies. Following Lykken's (1968, 1991) nomenclature for conducting replications in psychological research, if all of these longitudinal studies mirror one another, aligning their findings over multiple decades would constitute a series of *constructive replications*, which are the most scientifically compelling kind: The idea behind constructive replication is to vary systematically as many construct-irrelevant design features as possible over successive replications, while ensuring that the focal construct is preserved in each study. In Lykken's (1968) words,

> To obtain an ideal constructive replication, one would provide a competent investigator with *nothing more than* a clear statement of the empirical "fact" which the first author would claim to have established . . . and then let the replicator formulate his own methods of sampling, measurement, and data analysis . . . . We are interested in the *construct*, . . . not the *datum*. (p. 156)

In the current context, the studies we are aligning employed different measures, cohorts, time points, longitudinal intervals, investigators, and criteria. Yet, the focal construct, *spatial ability*, and its role in various educational and occupational pursuits remained the same. With this foundation, the following study was conducted.

Logic and Constructive Replication Sample

The specific objectives in this study were (a) to determine the extent to which spatial ability has operated consistently for decades in the prediction of educational and occupational criteria with particular emphasis on STEM domains, (b) to determine the extent to which early manifestations of exceptional spatial ability portend the development of STEM expertise, and (c) to demonstrate how neglect of this important dimension of cognitive functioning leads to untapped pools of talent for STEM domains.

The Shea et al. (2001) findings constitute, to our knowledge, the first demonstration that spatial ability adds incremental validity (beyond mathematical and verbal ability measures) in the prediction of educational–occupational criteria among talent search participants initially identified before age 13 on the basis of SAT-Math and SAT-Verbal scores. Some of their longitudinal outcomes, which include favorite and least favorite high school course (age 18 follow-up), college major (age 23 follow-up), and occupation (age 33 follow-up), are shown in Figure 2, as a function of their standing on these three abilities assessed at age 13 in standard deviation units. Mathematical ability is scaled on the x-axis, verbal ability on the y-axis, and spatial ability on the z-axis (notated by arrows in standard deviation units; arrows to the right are positive effect sizes for spatial ability, and arrows to the left are negative effect sizes for spatial ability). Essentially, this is a three-dimensional graph put in a two-dimensional representation. This figure will serve as a template for replication purposes. To visualize the location of each group in three-dimensional space, imagine the arrows to the right projecting outward (toward you) and the arrows to the left projecting inward (away from you), both perpendicular to the x- and y-axes; in this way, the psychological distance between these criterion groups can be pictured in the space defined by the three ability dimensions. Dotted lines are placed around the STEM groups to highlight their consistent pattern across all three time points. We predicted that these patterns also would be observed in Project TALENT participants, whose 11-year longitudinal follow-up was conducted before these SMPY participants were identified in the 1970s at age 13.

It is important to keep in mind that although the SMPY participants were identified as intellectually talented in early adolescence (top 0.5% for their age-group), their patterns of specific abilities are readily distinguished as a function of contrasting educational–occupational group membership. With respect to spatial ability, the focal construct under analysis here, the consistently above-average spatial ability of participants in STEM educational degree groupings and occupations reveals the importance of spatial ability in STEM arenas (as indicated by rightward-pointing arrows across all four panels of Figure 2). Within the dotted boxes in each panel of Figure 2 are the STEM groups. However, to clarify the graph, examine just the physical science group in Panel C. This group has a positive z-score value (relative to the other groups) on mathematical, verbal, and spatial ability. Stated differently, those individuals who majored in physical science had higher mathematical, verbal, and spatial abilities relative to those who majored in other areas. In contrast, examine the humanities group in Panel C. This group has a positive z-score value on verbal ability but a negative z-score value on both mathematical and spatial ability relative to the other groups. What this means is that those individuals who majored in the humanities had relatively higher verbal ability but relatively lower math and spatial ability in comparison to those with other majors.

Consistently lower levels of spatial ability, indicated by arrows pointing to the left, are associated with domains outside of STEM. For example, referring back to the physical science group in Panel C, this group's z score on spatial ability was 0.34 (the length of the rightward-pointing arrow), whereas the humanities group's z score on spatial ability was -0.34 (the length of the leftward-pointing arrow). This means that these two groups are 0.68 standard deviations apart on spatial ability, even though both groups were above the normative mean on spatial ability (Shea et al., 2001). Hence, relative strengths and weaknesses manifested during adolescence are related to contrasting outcomes in education and the world of work. Jointly, these successive panels demonstrate how spatial ability operates over the life span (after high school, after college, and at age 33), regardless of whether it is measured. That is, whereas mathematical and verbal ability measures were used to identify these participants and similar measures were subsequently used throughout their educational careers as selection tools, spatial ability was assessed experimentally only at the time of their initial identification; spatial ability was not then used and is very rarely currently used in educational selection for advanced degrees or professional careers. Spatial ability played a clear role for these intellectually talented youths in domains in which it is placed at a premium (as well as those in which it is not). Multiple examples of how spatial ability operated in the attainment of educational and

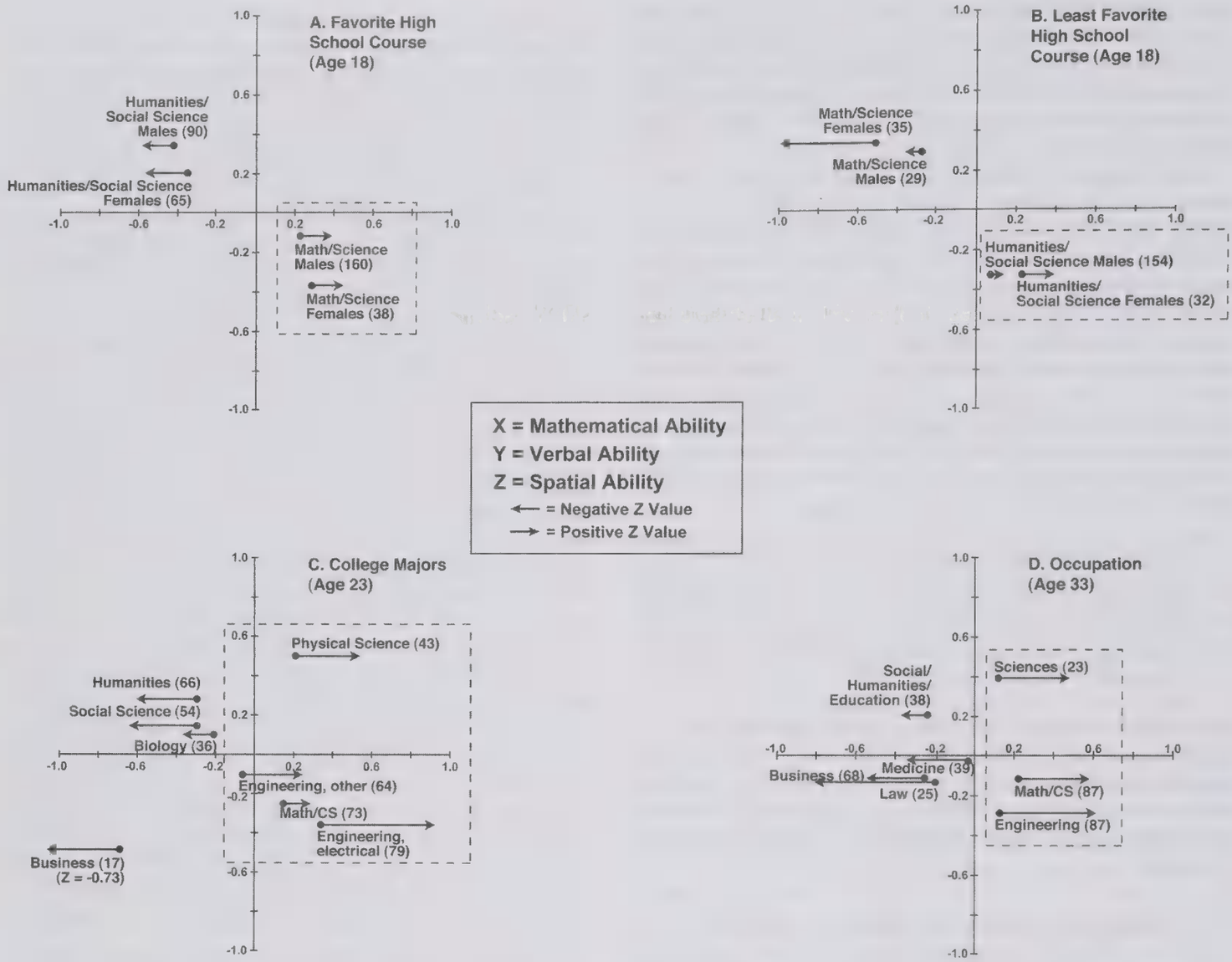


Figure 2. Shown are trivariate (X/Y/Z = Mathematical/Verbal/Spatial) means for (Panel A) favorite and (B) least favorite high school course at age 18, (C) college majors at age 23, and (D) occupation at age 33. Mathematical, verbal, and spatial ability are on the *x*-, *y*-, and *z*-axes respectively (arrows to the right indicate a positive *z* value; arrows to the left indicate a negative *z* value). Panels A and B are standardized within sex; Panels C and D are standardized across sexes. For Business in Panel C, note that the length of the arrow is actually $z = -0.73$. Figure adapted from Shea, Lubinski, & Benbow (2001). CS = computer science.

occupational criteria in STEM areas (where spatial ability is quite important) are highlighted by the rightward-pointing arrows for each of the STEM groups, which are contained within the dotted-line boxes in all four panels of Figure 2. The STEM groups were higher on spatial ability relative to the other groups; we refer to these groups later in the article. A key question is, has spatial ability been operating in this way in normative samples as well and, if so, for how long?

Replication Sample: Project TALENT

In this study, we formed a number of educational and occupational groupings using Project TALENT's 11-year follow-up data to reveal the extent to which spatial ability assessed in adolescence is a salient characteristic among individuals who subsequently go on to achieve educational and occupational credentials in STEM. In addition, to ascertain the extent to which findings uncovered with Project TALENT mirror those in the Shea et al. (2001) study, we scrutinized the pattern similarity of specific abilities in both data sets. The question here is, when calibrated against STEM criteria, will both data sets reveal a consistent pattern (i.e., high mathematical and spatial ability and relatively lower verbal ability over multiple longitudinal time frames)? Moreover, the sample sizes available in Project TALENT allow us to examine whether higher levels of spatial ability differentiate people operating at more advanced educational levels within STEM. Finally, we identify the proportion of participants in the top 1% of spatial ability who are not in the top 1% on either mathematical or verbal ability and, hence, are lost by identification procedures restricted to mathematical or verbal ability; we examine the

educational and occupational outcomes of these students to understand better what kinds of students modern talent search procedures are failing to identify for advanced educational opportunities. For further, more detailed reading on the methodological approach of creating criterion groups based on educational and occupational credentials and, subsequently, examining the salient characteristics among members of each group at earlier time points for clues about the psychological antecedents giving rise to them, see Dawis (1992); Dawis and Lofquist (1984); and especially Humphreys et al. (1993) and references therein.

Method

Participants and Measures

Participants were drawn from the Project TALENT data bank, an ideal sample for our purposes here due to its comprehensiveness, size, and longitudinal time frame. Project TALENT's initial data collection in 1960 consisted of a stratified random sample of the nation's high school population (Flanagan et al., 1962). Students in the 9th through 12th grades were assessed on a wide range of tests and questionnaires over a 1-week period; the entire sample included roughly 50,000 males and 50,000 females per grade level, for a total *N* of approximately 400,000. Included in the tests were a number of measures designed to assess cognitive abilities (e.g., mathematical, verbal, and spatial ability), as well as information tests (on content areas including art, biology, engineering, journalism, and physics) and measures of attitudes, interests, and personality traits. Participants also completed a 398-item questionnaire on their lives (e.g., topics such as family, school, work, hobbies, and health). Tests and questionnaires were administered over a period of 1 week. These materials can be obtained through the American Institutes for Research, Palo Alto, California (see Flanagan et al., 1962, and Wise et al., 1979, for a thorough description of the range of tests and questionnaires administered).

Longitudinal Data

Project TALENT includes longitudinal data taken 1, 5, and 11 years after graduation from high school (Wise et al., 1979). For this study, we examined the 11-year follow-up data and focused on those who reported their highest degree received (a bachelor's, master's, or doctoral degree) and occupation.

Research Design

The conceptual framework used to form our ability measures stems from the hierarchical organization of cognitive abilities (Carroll, 1993). A cogent simplification of Carroll's model is the radex organization or scaling of cognitive abilities (Snow, Corno, & Jackson, 1996; Snow & Lohman, 1989). The radex organizes cognitive abilities around three content domains: quantitative/numerical, spatial/pictorial, and verbal/linguistic (or mathematical, spatial, and verbal domains, respectively); the communality cutting across these three content domains distills the higher order construct of general intelligence (*g*). The latter denotes the sophistication of the intellectual repertoire. Figure 3 constitutes a visual representation of the radex, which is made up of an

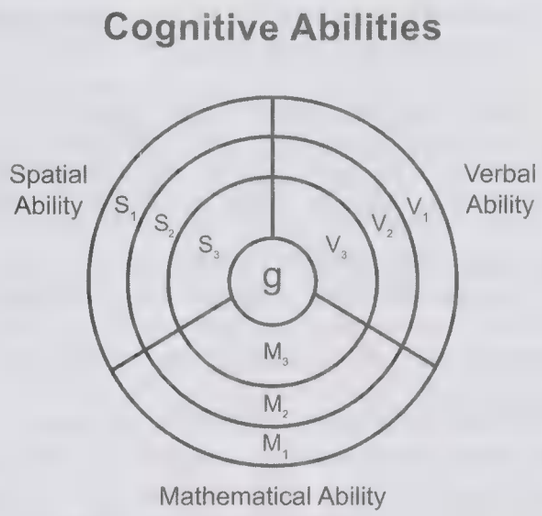


Figure 3. The radex of cognitive abilities organizes cognitive abilities around mathematical, spatial, and verbal content domains, with the higher order construct of *g*, or general intelligence, at the center representing the communality shared by these specific abilities.

infinite number of simplexes and circumplexes. An example of a simplex would be a continuum running from the centroid, or *g*, through S_3, S_2, and S_1, respectively. Along this simplex, the test content is spatial and the test complexity diminishes as one moves from the center to the periphery (e.g., a test located in S_3 would be more complex than a test located in S_1). Thus, for simplexes, test content remains comparable but complexity changes. An example of a circumplex is a circular band running through S_2, V_2, and M_2, respectively. Along this circumplex, the test content would vary, being spatial (reasoning with figures and shapes), verbal (reasoning with words), or mathematical (reasoning with numbers); however, the test complexity would remain comparable. Within the radex, tests varying in content and complexity can be found, and these two dimensions are necessary for locating a specific test in this space. The radex is a very efficient way of arranging the many different kinds of psychometric indicators of cognitive abilities. To the extent that measures covary with one another, they are close to one another in this two-dimensional space. Correspondingly, to the extent that measures do not covary with one another, they are distant from one another in this space (cf. Lubinski & Dawis, 1992, p. 8, for an empirical example). Thus, spatial ability, the focal construct under investigation, is distinguished from the more familiar constructs of mathematical and verbal ability in the context of a hierarchical illustration of the radex organization of cognitive abilities (see Figure 3).

Ability composites. We formed three ability composites with which to measure the three components found in the radex (Snow & Lohman, 1989; Wise et al., 1979): mathematical, spatial, and verbal ability. The Mathematical Composite consisted of four tests:

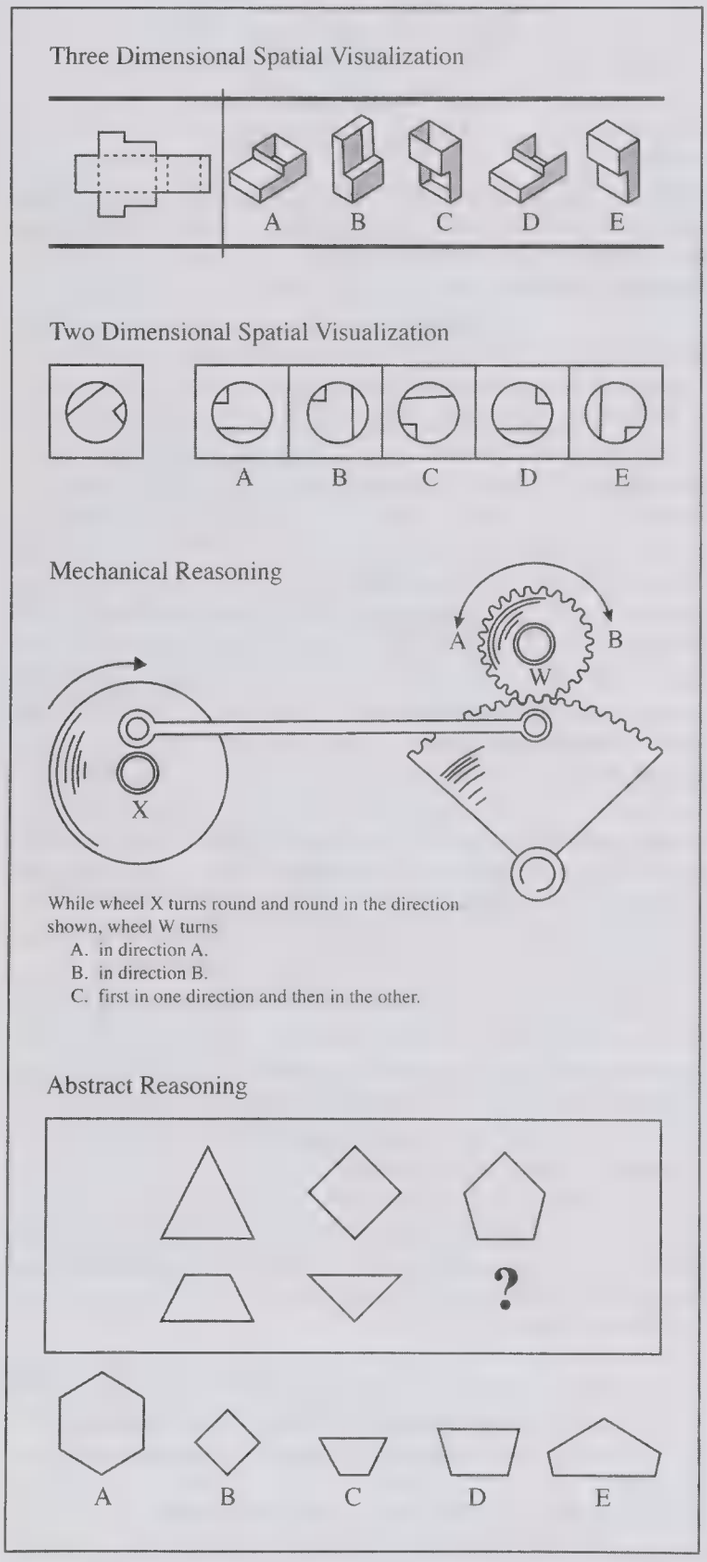
1. Mathematics Information (23 items measuring knowledge of math definitions and notation). A sample item might be "Which of these is an irrational number?"

2. Arithmetic Reasoning (16 items measuring the reasoning ability needed to solve basic arithmetic items). A sample

item might be "A man pays 4% sales tax on a chair. The tax is $6.00. How much did the chair cost?"

3. Introductory Mathematics (24 items measuring all forms of math knowledge taught through the 9th grade). A sample item might be "Suppose the sum of 2 two-digit numbers is a three-digit number. What is the first digit of the sum?"

4. Advanced Mathematics (14 items covering algebra, plane and solid geometry, probability, logic, logarithms, and basic calculus). A sample item might be "Which of these equations has no real roots?"



To maximize construct validity (see below), we assigned the following weights based on scale variances and covariances to these constituents: Mathematical Composite = 0.55 × [Mathematical Information] + 1.0 × [Arithmetic Reasoning] + 0.55 × [Introductory Mathematics] + 1.0 × [Advanced Mathematics].

The Verbal Composite was composed of three measures:

1. Vocabulary (30 items that measure general knowledge of words). A sample item might be "Placate means" with answer choices following.

2. English Composite (113 items measuring capitalization, punctuation, spelling, usage, and effective expression). A sample item for covering usage might be "He _____ ready yet; A. isn't, B. ain't, or C. aren't."

3. Reading Comprehension (48 items measuring the comprehension of written text covering a broad range of topics). A sample item in this section would be similar to a typical reading comprehension item found on an exam such as the SAT.

Verbal Composite = 2.5 × [Vocabulary] + 1.0 × [English Composite] + 1.25 × [Reading Comprehension].

Finally, the Spatial Composite was composed of four measures (and because the focus of this study is on spatial ability, item types for each are illustrated in Figure 4):

1. Three-Dimensional Spatial Visualization (16 items measuring the ability to visualize two-dimensional fig-

Figure 4. Three-dimensional spatial visualization. Each problem in this test has a drawing of a flat piece of metal at the left. At the right are shown five objects, only one of which might be made by folding the flat piece of metal along the dotted line. You are to pick out the one of these five objects which shows just how the piece of flat metal will look when it is folded at the dotted lines. When it is folded, no piece of metal overlaps any other piece or is enclosed inside the object.

Two-dimensional spatial visualization. In this test each problem has one drawing at the left and five similar drawings to the right of it, but only one of the five drawings on the right exactly matches the drawing at the left if you turn it around. The rest of the drawings are backwards even when they are turned around. For each problem in this test, choose the one drawing which, when turned around or rotated, is exactly like the basic drawing at the left.

Mechanical reasoning. This is a test of your ability to understand mechanical ideas. You will have some diagrams or pictures with questions about them. For each problem, read the question, study the picture above it, and mark the letter of the answer on your answer sheet.

Abstract reasoning. Each item in this test consists of a set of figures arranged in a pattern, formed according to certain rules. In each problem you are to decide what figure belongs where the question mark is in the pattern. To do this you have to figure out what the rule is according to which the drawings change, going from row to row, and what the rule is for the changes going from column to column. The items have different kinds of patterns and different rules by which the drawings change. The question mark in the lower right corner of each box shows where a figure is missing in the pattern. You are to decide which of the five figures (A, B, C, D, or E) under the pattern belongs where the question mark is.

ures after they had been folded into three-dimensional figures).

2. Two-Dimensional Spatial Visualization (24 items measuring the ability to visualize two-dimensional figures when they were rotated or flipped in a plane).

3. Mechanical Reasoning (20 items measuring the ability to deduce relationships between gears, pulleys, and springs as well as knowledge of the effects of basic physical forces, such as gravity).

4. Abstract Reasoning (15 items constituting a nonverbal measure of finding logical relationships in sophisticated figure patterns).

Spatial Composite = 3.0 × [3-D Spatial Visualization] + 1.0 × [2-D Spatial Visualization] + 1.5 × [Mechanical Reasoning] + 2.0 × [Abstract Reasoning].

The above weights were derived by Humphreys to form composites that mirror the SAT-M and SAT-V and the location of spatial ability within the context of the hierarchical organization of cognitive abilities (see the radex in Figure 3), and these composites have been used extensively in other research (Gohm et al., 1998; Humphreys et al., 1993; Lubinski & Humphreys, 1990a, 1990b, 1996). The intercorrelations of these composites for the 9th-grade cohort were .61, .59, and .76 for mathematical–spatial, verbal–spatial, and mathematical–verbal, respectively. In the current study, we added the English Composite and Advanced Mathematics to their respective composites initially derived by Humphreys to augment the ceiling of each scale, but this modification changed the intercorrelations of these three composites by an average of only .01 correlational units for each sex across all four cohorts. Hence, their conceptual equivalency and empirical interchangeability were preserved. Humphreys has estimated that the reliabilities of these, or very similar, composites are approximately .90 (Humphreys, 1991). These estimates were based on conservative estimates of parallel form reliabilities of the components.

Design. Participants were included if they had complete ability data at Time 1: (*n*s: 9th grade, males = 47,440, females = 47,496; 10th grade, males = 46,112, females = 45,199; 11th grade, males = 41,766, females = 43,751; 12th grade, males = 36,375, females = 38,526). Would the educational and occupational group membership of these participants, assessed 11 years after their high school graduation, retrospectively isolate distinctive ability profiles based on their adolescent assessments? If so, and if these findings mirrored those uncovered in Shea et al. (2001) over a 20-year interval and nonoverlapping time frame (see Figure 2), these corresponding function forms would constitute a *constructive replication* (Lykken, 1968, 1991).

In this context, it is worth mentioning Meehl's (1978) point that in the early stages of theory construction, function form is often more important than statistical significance (see also Steen, 1988). For example, the patterns of specific abilities in all four panels of Figure 2 reveal consistent function forms. Across all four panels, meaningful STEM outcomes are found in roughly the same location as a function of the three specific abilities; over all time points, they reveal the same pattern (see the groups in the dotted-line boxes). That is, the same function form or pattern of specific abilities distinguishes the STEM groups from the other criterion groupings over these multiple time points (high school, college, and occupation). The precise location of the points on each panel is not as critical as the overall pattern formed by the specific abilities over time. They appear to be operating in the same way, and the pattern maintains its function form. We hypothesized that similar function forms would be found using Project TALENT (a different cohort and nonoverlapping time frame) and that these patterns would be of sufficient magnitude to be of substantive interest to psychological practitioners, applied researchers, and theoreticians interested in educational readiness and adult achievement. Finally, although both function form and statistical significance are evaluated here, the former is more central because, given the sample sizes, virtually all group contrasts will manifest statistically significant differences on the specific abilities under analysis.

We made the following hypotheses:

1. The pattern or function form uncovered from the participants in SMPY on the three specific abilities will be mirrored by those in Project TALENT when calibrated against conceptually meaningful educational and occupational criterion groupings.

2. The importance of spatial ability will increase as a function of more conceptually demanding STEM criteria (e.g., advanced educational degrees in STEM: bachelors, masters, and doctorates).

3. To the extent possible, findings taken from the Graduate Record Exam will mirror those of Project TALENT and SMPY.

4. An appreciable percentage of young adolescents with talent for STEM and other domains in which spatial ability is placed at a premium are missed by contemporary talent searches and current selection procedures for STEM careers.

To begin to examine this series of hypotheses, after selecting participants in each cohort as a function of complete ability and group membership data, we computed *z* scores for all three specific abilities within cohort and, then, over all four cohorts. Data for each criterion group (within highest degree and occupation) were aggregated. We then plotted each group's mean *z* score for all three specific abilities (see Figure 5). Appendix A includes the respective sample sizes broken down by sex for each degree and occupation included in the groups plotted over all four panels of Figure 5.[1]

[1] We conducted analyses for males and females separately within each grade level (9th, 10th, 11th, and 12th) and found that the pattern was strikingly similar.

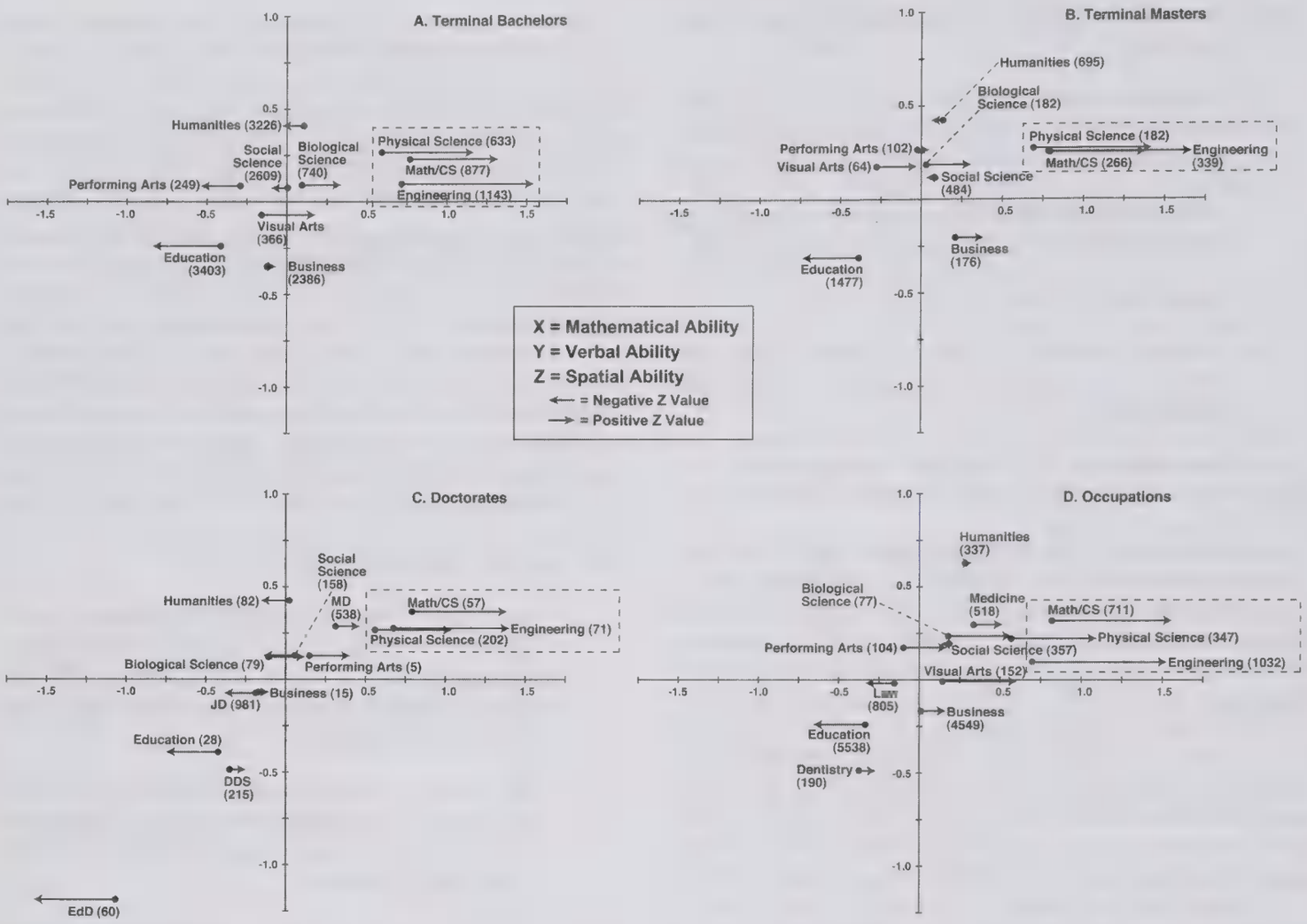


Figure 5. Trivariate means for (Panel A) bachelors, (B) masters, (C) doctorates, and (D) occupations of those individuals whose data were included in Panels A, B, and C. Panels A through D are standardized across sexes. Mathematical ability is on the *x*-axis, and verbal ability is on the *y*-axis; an arrow from each group mean indicates either positive (to the right) or negative (to the left) spatial ability. Breakdowns by sex are reported in Appendix B. The standard error of the mean for $n = 500$ was 0.04 and for $n = 1{,}000$ was 0.03. Data are from Project TALENT. CS= computer science.

Results

Ability Pattern

A number of the groups included in Figure 5 are conceptually similar to those in the Shea et al. (2001) panels found in Figure 2. Those that are not were included to provide a more detailed context for how these ability composites operate within a more broadly defined educational–occupational criterion space. Figure 5 includes those individuals who reported at the 11-year follow-up that their highest degree received was a BA or BS (Panel A; males = 8,446, females = 7,186), a MA or MS (Panel B; males = 2,383, females = 1,584), or a doctorate (Panel C; males = 2,293, females = 198). In addition to PhDs, which correspond to the degrees found in Panels A and B, Panel C contains MDs, JDs, DDSs, and EdDs for comprehensiveness. Panels A, B, and C included nonoverlapping groups of participants, as only highest degrees received were plotted. Panel D (males = 10,389, females = 4,328) included all participants who reported a degree as shown in Panels A, B, and C and who also reported an occupation 11 years after high school graduation. Each graph parallels the Figure 2 template; mathematical (*x*-axis), verbal (*y*-axis), and spatial (*z*-axis) abilities are plotted in standard deviation units. Sample sizes are next to each group in parentheses.

One can see in each panel a general ability, or *g*, gradient (driven by the communality running through spatial, mathematical, and verbal ability), extending from the lower left quadrant to the upper right quadrant. It is also evident that the patterns of specific abilities (within each group and across groups) shown in the four panels are strikingly similar. As an example, within the dotted-line boxes in each panel of Figure 5 are the STEM groups (as in Figure 2). Examination of the physical science group in Panel C shows that this group has a positive *z*-score value (relative to the mean) on mathematical, verbal, and spatial ability. It is important here to note that in both Figures 2 and 5 the STEM groups all have rightward-pointing arrows, which indicate higher spatial ability

(relative to the other groups in each panel). The rightward-pointing arrow in Panel C for the physical sciences is 0.45 standard deviation units greater than the mean, whereas the humanities grouping is −0.15 standard deviation units below the mean; therefore, these two groups are 0.60 standard deviations apart on spatial ability. This constitutes one of many examples of an important constructive replication of function form or pattern across Figures 2 and 5 for spatial ability, which is central to our thesis.[2] It is important to note here that the locations of mathematical (*x*-axis) and verbal (*y*-axis) abilities for some groups are offset somewhat over Panels A through D in Figure 5 in comparison to those of Figure 2, but for good reason.

Dawes (1975) and Sackett and Yang (2000) have discussed how structural relationships among measures can change when samples under analysis are selected with predictors that go into subsequent analyses.[3] Participants summarized in Figure 2 were selected using the SAT, whereas Figure 5 participants were a stratified random sample of the nation's high school population who subsequently went on to earn advanced educational credentials. As noted before, neither cohort was selected on spatial ability. It is striking, therefore, how clear-cut the findings are for corresponding groups on spatial ability, especially in STEM domains. Moreover, other domains, such as biology and the visual arts, also appear to draw on spatial ability. For other, more general consistencies found in these data, Appendix B provides the overall level of general and specific abilities that these groups manifested at Time 1.[4]

In Figure 6, we extend these analyses in part to another contemporary sample by plotting bivariate (X/Y = Mathematical/Verbal) means for the mathematical and verbal ability composites from Project TALENT for participants who later went on to secure graduate degrees (black circles). The bivariate means for these participants are connected with lines to the bivariate means of contemporary graduate students on the basis of corresponding mathematical and verbal measures (white circles) on the Graduate Record Examination (GRE). Figure 6 is similar to Figures 2 and 5 in that it includes mathematical ability on the *x*-axis and verbal ability on the *y*-axis, but it is different in that it does not include spatial ability (rightward- and leftward-pointing arrows), inasmuch as the GRE does not include a spatial measure. Again, mathematical ability is a salient attribute of students seeking to develop STEM expertise. As described in the caption, each GRE grouping represents thousands of prospective graduate students. Although the GRE does not assess spatial ability, given the consistencies between the GRE and Project TALENT's mathematical and verbal ability scales and the well-known longitudinal consistencies of the covariance structures between measures of these constructs (Carroll, 1993; Johnson & Bouchard, 2007a, 2007b; Lubinski, 2004; Snow et al., 1996), it would be surprising if modern spatial ability assessments did not uncover patterns consistent with the other longitudinal findings (see, e.g., Figures 2 and 5). Essentially, we anticipated that the spatial ability arrows on the *z*-axis for the GRE data, if plotted, would reflect those found in Figures 2 and 5. In Figure 6, graduate degrees in the humanities are high on the *y*-axis, and a salient cluster of graduate degrees in STEM are located far to the right on the *x*-axis (engineering, math/computer science, and physical science); business and education also demonstrate a co-occurrence of location across both data sets in the space defined by these dimensions. Consistent locations are therefore found over a 40-year interval.

Spatial Ability Level

With respect to overall level of ability, the likelihood of earning an advanced degree in STEM as a function of spatial ability is depicted in Figure 7. Using 11-year follow-up data from Project TALENT, we classified the subset of participants with STEM degrees into three groups (as a function of their highest terminal degree): bachelor's, master's, or PhD. This was done within each cohort separately, and then findings from all four cohorts were aggregated. Finally, we plotted the proportion of each degree within each stanine based on spatial ability stanine in high school. It becomes clear from these findings that spatial ability plays an important role in achieving advanced educational credentials in STEM. From an epidemiological point of view (Lubinski & Humphreys, 1996, 1997), the likelihood or promise of earning an advanced degree in STEM areas increases as a function of spatial ability. These findings are clear: 45% of all those holding STEM PhDs were in Stanine 9 (or within the top 4%) on spatial ability 11+ years earlier, and nearly 90% were in Stanine 7 or above. That is, less than 10% of those holding STEM PhDs were below the top quartile in spatial ability during adolescence. In comparison to the 45% of STEM PhDs in Stanine 9, for example, about 30% of those holding STEM terminal master's degrees and 25% of those holding STEM terminal bachelor's degrees were in Stanine 9, or the top 4% of spatial ability. We can conclude that the importance of spatial ability for STEM increases as a function of successively more advanced educational credentials.

Finally, is there a way to determine the extent to which modern talent searches miss high-potential students gifted in spatial ability? Some summer residential programs for talented youths require scores in the top 1% on either verbal or mathematical ability measures to ensure readiness to take advantage of the fast-paced learning demands

---

[2] For all four panels in Figure 5, we also conducted the following series of incremental validity analyses. For each of the three terminal degrees and occupations, we dummy coded each STEM cluster as 1 and the remainder of the groups as 0 (and we utilized this as a criterion variable). We then ran multiple regression analyses for each panel by first entering mathematical ability and verbal ability and then determining the incremental validity, or multiple-R^2 increment, for spatial ability in predicting this dichotomous variable (STEM, non-STEM). The incremental validity of spatial ability over mathematical ability and verbal ability for all four panels was statistically significant, as anticipated, and the multiple-R^2 increment averaged .04 (accounting for an additional 4% of criterion variance).

[3] The relationship found in nature between two variables actually can be inverted (a positive covariance can become negative) when selection occurs on a third variable. For example, among undergraduates applying to graduate school, their composite Graduate Record Exam (GRE) score is positively correlated with their undergraduate grade point average (GPA) but is negatively correlated among those selected within a particular school. The reason is that a low GRE composite can be compensated for by a high GPA, and the inverse is true for low GPAs. But for graduate school, the low-GRE, low-GPA students tend not to be selected; this removal of the southwest quadrant of the fourfold table (GRE/GPA, High/Low) switches a positive covariance to a negative one.

[4] In a related vein, as the 9th grade of Project TALENT is closest to the SMPY talent search population in time of initial testing, we conducted an analysis to determine how similar all four cohorts were to the 9th-grade cohort alone. The average difference between all four grades and the 9th-grade sample for the respective correlations for mathematical, spatial, and verbal abilities was less than 0.03 correlational units.

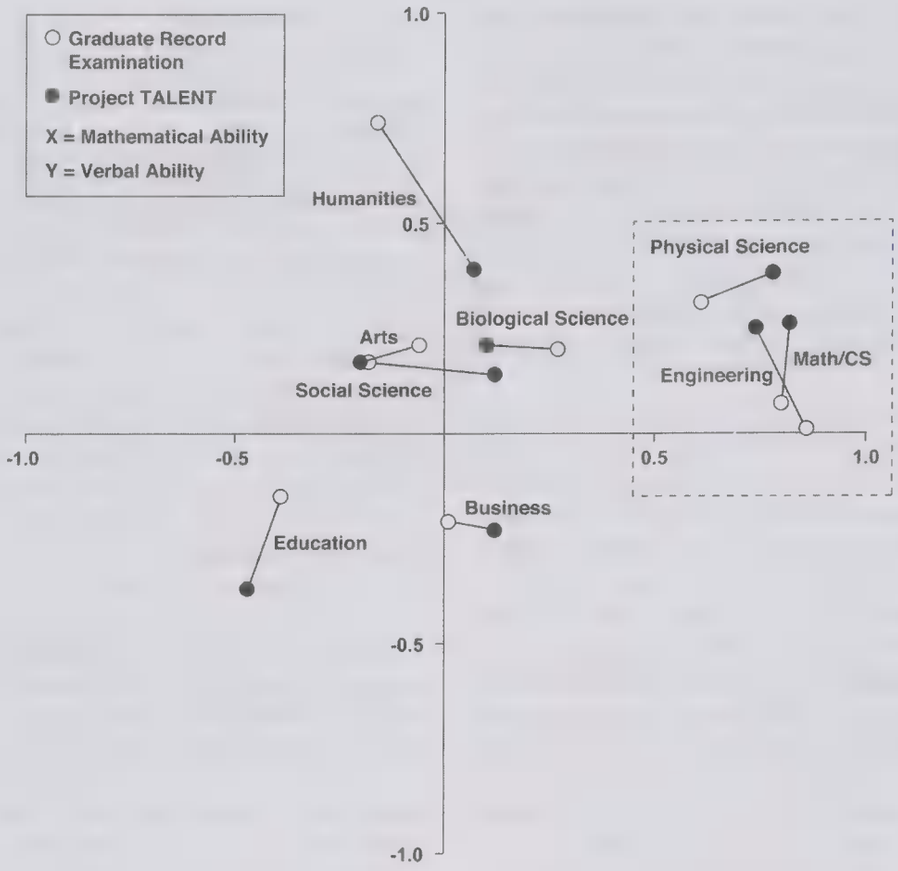


Figure 6. Data on the Graduate Record Examination (GRE) taken on individuals in the respective educational groups tested between July 1, 2002, and June 30, 2005, are graphed alongside commensurate data from Project TALENT. For each group, *z* scores on mathematical (*x*-axis) and verbal (*y*-axis) abilities respectively are plotted (standardized within the groups represented). White circles = GRE data. Black circles = Project TALENT data. A line was drawn connecting the two data points for each group to illustrate the distances between points of the same field. The *z* scores for each group were computed by taking the difference between the group mean and the overall mean for each subtest and dividing by the population standard deviation of that subtest. The total number of those taking the GRE for each subtest for these data was 1,245,872 for GRE-Mathematical (GRE-M) and 1,245,878 for GRE-Verbal (GRE-V). The respective groups were chosen to mirror the ones in Figure 2 and were as follows (with *n*s for GRE-V and GRE-M, respectively): engineering (56,368, 56,374); physical science (22,487, 22,485): chemistry, earth, atmospheric, and marine sciences, and physics and astronomy; math/computer science (33,107, 33,108): computer and information sciences, mathematical sciences; biological science (37,579, 37,576); humanities (37,468, 37,435): English language and literature, foreign languages and literatures, history, philosophy, and religion and theory; social science (101,085, 101,064); arts (20,040, 20,057): architecture and environmental design, art history, theory, and criticism and arts, performance and studio; business (8,357, 8,357); education (43,844, 43,835). Project TALENT data (PT-M, PT-V) were analyzed within MAs, MSs, and PhDs specifically to best mirror the GRE data. Correlations between the means for the respective educational groups were computed between GRE-M and PT-M ($r = .93$, $p < .01$), GRE-V and PT-V ($r = .77$, $p < .05$), and GRE-M + V and PT-M + V ($r = .96$, $p < .01$). The average difference across all three methods of comparison (i.e., correlations GRE-M minus PT-M, GRE-V minus PT-V, and GRE-M + V minus PT-M + V) and major groupings was less than the absolute value of 0.04, 0.02, and 0.02, respectively. The standard error of the mean for $n = 500$ was 0.04 and for $n = 1{,}000$ was 0.03. GRE data were taken from http://www.ets.org/Media/Tests/GRE/pdf/5_01738_table_4.pdf and http://www.ets.org/Media/Tests/GRE/pdf/4_01738_table_1a.pdf

in their programs, and some even require scores in the top 0.5% (Benbow & Stanley, 1996; Colangelo et al., 2004). Thus, there exists another question that Project TALENT can answer: How many spatially gifted students are missed for such programs by current talent search practices, which focus only on mathematically and verbally talented youths? Within the three ability composites assembled for this study, 70% of the top 1% in spatial ability did not make the cut for the top 1% on either the math or the verbal composite; yet, these individuals are highly talented in spatial ability. Figure 8 presents data on the educational and occupational outcomes of this 70% in terms of their credentials in STEM domains (top panel) and the visual arts (bottom panel). The latter group was added to highlight the longstanding recognition of the importance of spatial ability for many of the creative arts. The black bars show the base rates for these outcomes in Project TALENT; the overall bars (black + gray) represent those in the top 1% on the Spatial Composite who were not in the top

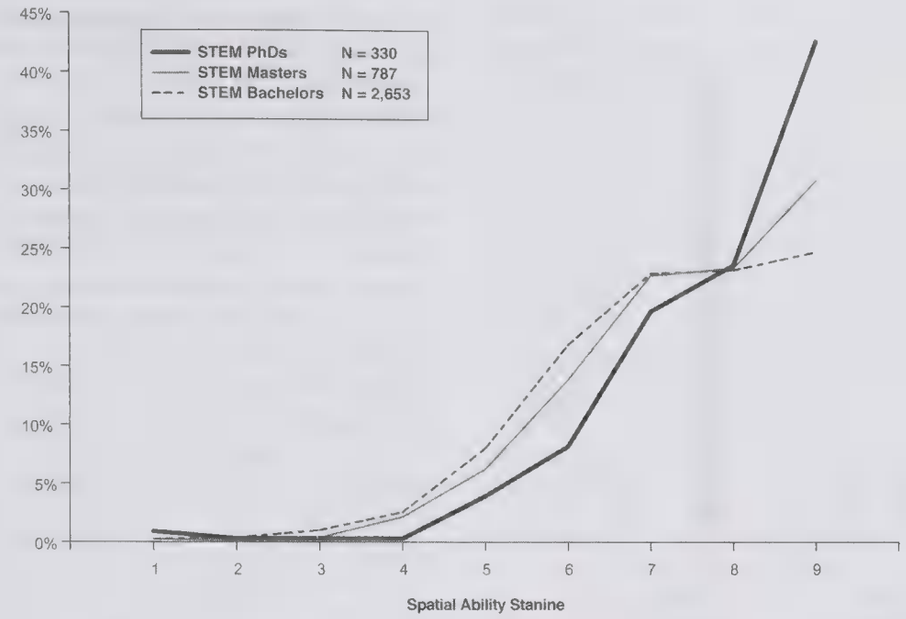


Figure 7. This figure includes the proportion of each degree group (bachelors, masters, and PHDs) as a function of spatial ability. Along the *x*-axis are the spatial ability stanines (numbered 1 through 9). STEM = science, technology, engineering, and mathematics.

1% on either the Math or Verbal Composites. This potential, currently being missed, constitutes a rather sizable pool of untapped talent. Among those in the top 1% in spatial ability but not in the top 1% in mathematical or verbal ability, a large proportion earned STEM and visual arts degrees and entered STEM and visual arts occupations well beyond base rate expectations.

Discussion

Longitudinal findings uncovered in this study combined with results of earlier investigations (Super & Bachrach, 1957) and recent longitudinal findings on intellectually precocious youths (Shea et al., 2001; Webb et al., 2007) suggest at least three generalizations: First, spatial ability is a salient psychological characteristic among adolescents who subsequently go on to achieve advanced educational and occupational credentials in STEM. Second, spatial ability plays a critical role in structuring educational and occupational outcomes in the general population as well as among intellectually talented individuals. Third, contemporary talent searches miss many intellectually talented students by restricting selection criteria to mathematical and verbal ability measures.[5]

Given the body of evidence now available and the fresh empirical findings presented here on thousands of high school students tracked 11 years following their high school graduation, sufficient support has accrued to demonstrate that the importance of spatial ability in STEM domains has been operating for several decades. Just as F. L. Schmidt and Hunter (1998) concluded in their 85-year review of the role that general intelligence plays in the world of work ("more research is not needed"), we conclude that enough empirical evidence has accrued to register another rare example of a solid empirical generalization within the human psychological sciences. This does not mean, however, that other research is not needed. The kind of research that is needed now is in how to utilize spatial ability for student selection, instruction, and curriculum design and in how to refine educational interventions and procedures on the basis of individual differences in spatial ability (Corno et al., 2002; Lubinski, 2004, pp. 105–106).

In addition, measures of spatial ability should be incorporated into models of educational and occupational development to ascertain the role spatial ability plays relative to other abilities and relevant nonintellectual determinants (Lubinski & Benbow, 2000, 2006). Given the evidence presented here, psychological modeling of STEM outcomes must incorporate spatial ability to avoid being incomplete or underdetermined (Lubinski, 2000; Lubinski & Humphreys, 1997). This is particularly true among those who go on to develop especially high levels of STEM expertise (cf. Figures 2, 5, 7, and Appendix B).

Furthermore, expanding admissions criteria for talent searches currently focused on identifying intellectually talented youths

[5] There have been some discussions in visible outlets and based on very small samples that socioeconomic status (SES) moderates the sex difference in spatial ability (Levine, Vasilyeva, Lourenco, Newcombe, & Huttenlocher, 2005). Levine et al. has been cited by a number of recent investigations (Alexander & Son, 2007; Bergemann et al., 2008; Chabris & Glickman, 2006; Ehrlich, Levine, & Goldin-Meadow, 2006; Hackman & Farah, 2009; Newcombe & Uttal, 2006; Noble, McCandliss, & Farah, 2007; Penner & Paret, 2008; Silverman, Choi, & Peters, 2007) as documenting this relationship. As Newcombe and Uttal (2006) further generalize, "We need to delineate why and how some of the core abilities that all humans have come to be developed to different degrees in ways that depend on interactions of SES and gender" (p. 395). We conducted an analysis with our spatial ability composite (see Appendix C), in which we divided the Project TALENT SES variable into four quartiles and examined by cohort and by sex at each level of SES the hypothesis that SES moderates the sex difference in spatial ability ($n \approx 10{,}000$ in each cell). As can be seen in Appendix C, it is simply not the case that SES moderates the sex difference in spatial ability.

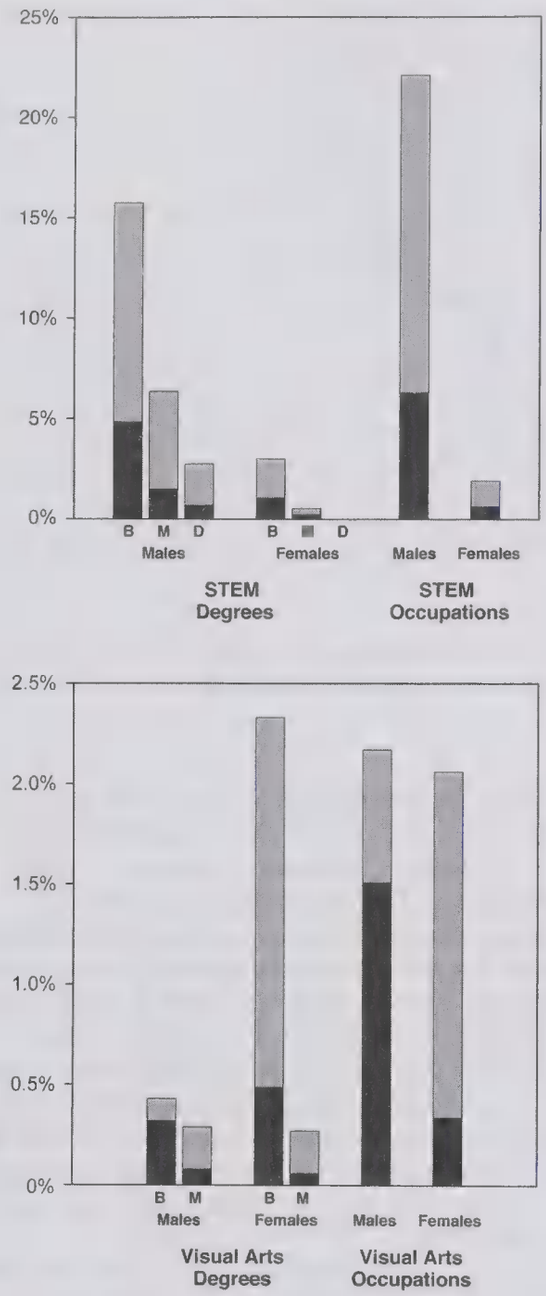


Figure 8. The top panel includes (left) the proportion of the top 1% in spatial ability who were not in the top 1% in mathematical or verbal ability who earned STEM degrees and (right) occupations broken down by males and females, respectively. The bottom panel includes the proportion of this population who earned visual arts degrees and worked in related occupations. The black bars indicate the base rate in Project TALENT for the respective grouping. B = bachelor's degrees; M = master's degrees; D = doctorate degrees; STEM = science, technology, engineering, and mathematics.

solely on the basis of scores on mathematical and verbal measures should be considered, as Snow (1999) suggested. For example, in Project TALENT, over half of participants in the top 1% on the Spatial Composite were below the top 3% cut on both the Mathematical and Verbal Composites, and, thus, they would not be invited to participate in modern talent searches. Moreover, there is reason to believe that the educational needs of spatially talented youths are more unmet than those of mathematically or verbally talented youths, because the typical middle and high school curriculum has many more opportunities for developing mathematical and verbal ability than spatial ability (Colangelo et al., 2004; Lohman, 2005).[6]

Ackerman (1996; Ackerman & Heggestad, 1997) has shown that all three of the specific abilities examined here (mathematical, spatial, and verbal) have distinctive external correlational profiles with respect to conventional measures of interests, personality, and values. For example, verbal ability tends to covary positively with interests in the humanities and helping people and to covary negatively with interests in engineering and technical pursuits. The opposite is true for spatial ability. An examination of the intercorrelations of Project TALENT's ability and interest measures reveals that these trait clusters, too, have been observed for decades in normative samples (Shaycoft, 1967) and, thus, must be seen as relatively stable. Moreover, these patterns of covariation have been replicated with intellectually talented youths (D. B. Schmidt, Lubinski, & Benbow, 1998) and have emerged as salient weights in discriminant function analyses in the prediction of STEM criteria (Achter, Lubinski, Benbow, & Eftekhari-Sanjani, 1999; Wai, Lubinski, & Benbow, 2005; Webb et al., 2007). Therefore, the motivational proclivities for students selected on the basis of mathematical versus spatial versus verbal ability should be expected to differ. That is, intellectually talented students selected by extreme cutting scores on measures of mathematical versus spatial versus verbal ability should be expected to have different interest patterns as well as differential preferences for linguistic, quantitative, and nonverbal ideation or contrasting modes of learning and thought (Corno et al., 2002).

Just as mathematically and verbally talented students have profited for decades by talent searches that identify students especially able at verbal and mathematical reasoning and the provision of tailored, developmentally appropriate curriculum aligned to their precocious rates of learning (or reasoning with linguistic and numerical symbols, respectively), students talented in spatial ability are likely to profit from identification procedures utilizing measures of spatial ability followed by opportunities for developmentally appropriate curriculum involving their preferred mode of thought (reasoning with forms or shapes). Experimentation with accelerative and rigorous learning opportunities in architecture, engineering, robotics, and the physical sciences appear to be particularly warranted in order to nurture their form of talent.[7]

Finally, sex differences in relative levels of interests are important to take into consideration. Although the covariance structure of specific abilities and interests is comparable for males and for females, the sexes display mean differences in a number of interests; for instance, spatially talented females tend to be more interested in artistic pursuits than are spatially talented males and

[6] For further and more detailed reading on measures of spatial ability and their conceptual underpinnings, see Corno et al. (2002); Eliot (1987); Eliot and Smith (1983); Lohman (1988, 1994a, 1994b, 1996, 2005); and Vandenberg and Kuse (1978). For more historical accounts, which have acknowledged the importance of spatial ability for technical trades and professions, see Paterson, Elliott, Anderson, and Toops (1930); Smith (1964); and Vernon (1961).

[7] This move would also foster conditions for adding value to longitudinal models of creativity currently restricted to mathematical and verbal reasoning abilities (Park, Lubinski, & Benbow, 2007, 2008).

the inverse is true for engineering and mechanical activities (Lubinski & Benbow, 2006; D. B. Schmidt et al., 1998). These mean sex differences in interests correspond to findings, shown in Figure 8, that spatially talented females were more likely than similarly talented males to pursue artistic domains. These proclivities can and do change over time, but relative levels of interests (and competing interests) are always important to take into account (Geary, 1998, 2005; Gottfredson, 2003, 2005).

Cumulative Psychological Knowledge

Collectively, the findings reported here, when combined with Super and Bachrach's (1957) NSF report and linked to modern research on talent search participants (Shea et al., 2001; Webb et al., 2007), tell a cohesive story about the longitudinal stability of spatial ability and its psychological import (see Figures 2, 5, 7, and Appendix B). For decades, spatial ability has emerged as a salient psychological characteristic among young adolescents who go on to develop expertise in STEM domains (see Figure 7).

This fact is important for more general considerations, because in psychology the lack of cumulative knowledge upon which to build theory and practice is often bemoaned. Cronbach (1975) has discussed the short "half-life" of empirical generalizations in the social sciences (i.e., how quickly they decay) and wrote, "The trouble, as I see it, is that we cannot store up generalizations and constructs for ultimate assembly into a network" (p. 123). Similarly, Meehl (1978) has observed that the "soft areas of psychology lack the cumulative character of scientific knowledge" (p. 806). Leaders in industrial (Dunnette, 1966) and clinical psychology (Dawes, 1994) have echoed these remarks. The current study offers an example of how the human psychological sciences can generate cumulative knowledge. Teaming constructive replication with longitudinal inquiry appears to be a compelling way to achieve cumulative psychological knowledge by revealing consistent function forms both across and within cohorts over protracted intervals.

Conclusion

As I. M. Smith (1964) stated so well 45 years ago,

> The qualities which make for greatness in scientists and engineers are of a different kind; ability to think abstractly and analytically together with skill in visualizing spatial relations in two or three dimensions, . . . All these qualities, which are vitally important in almost all branches of science and engineering, are measured by appropriate tests of spatial ability. (p. 300)

Spatial ability's robust influence on STEM domains has been supported in this article through the presentation of findings that link decades of longitudinal research. Collectively, the studies presented here constitute a series of constructive replications revealing similarities in function form and pattern across time (Meehl, 1978, 1990; Steen, 1988); therefore, an empirical generalization may be ventured on the importance of spatial ability in scientific and technical domains. In addition, individuals who are high in spatial ability but not as exceptional in mathematical or verbal abilities constitute an untapped pool of talent for STEM domains. Currently, more research is needed on how to effectively structure educational opportunities to serve students talented in spatial ability. Such efforts, if successful, will contribute to the urgent social need of effectively identifying and developing scientific and technical talent for the information age.

References

Achter, J. A., Lubinski, D., Benbow, C. P., & Eftekhari-Sanjani, H. (1999). Assessing vocational preferences among gifted adolescents adds incremental validity to abilities: A discriminant analysis of educational outcomes over a 10-year interval. *Journal of Educational Psychology, 91,* 777–786.

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22,* 227–257.

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121,* 218–245.

Alexander, G. M., & Son, T. (2007). Androgens and eye movements in women and men during a test of mental rotation ability. *Hormones and Behavior, 52,* 197–204.

American Competitiveness Initiative. (2006). *American Competitiveness Initiative: Leading the world in innovation.* Washington, DC: Domestic Policy Council Office of Science and Technology.

Benbow, C. P., & Stanley, J. C. (1996). Inequity in equity: How "equity" can lead to inequity for high-potential students. *Psychology, Public Policy, and Law, 2,* 249–292.

Bergemann, N., Parzer, P., Kaiser, D., Maier-Braunleder, S., Mundt, C., & Klier, C. (2008). Testosterone and gonadotropins but not estrogen associated with spatial ability in women suffering from schizophrenia: A double-blind, placebo-controlled study. *Psychoneuroendocrinology, 33,* 507–516.

Bingham, W. V. (1937). *Aptitudes and aptitude testing.* New York: Harper.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, England: Cambridge University Press.

Chabris, C. F., & Glickman, M. E. (2006). Sex differences in intellectual performance: Analysis of a large cohort of competitive chess players. *Psychological Science, 17,* 1009–1107.

Colangelo, N., Assouline, S. G., & Gross, M. U. M. (Eds.). (2004). *A nation deceived: How schools hold back America's brightest students.* Iowa City: University of Iowa.

Corno, L., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, D., Porteus, A. W., et al. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow.* Mahwah, NJ: Erlbaum.

Cronbach, L. J. (1957). Two disciplines of scientific psychology. *American Psychologist, 12,* 671–684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30,* 116–127.

Dawes, R. M. (1975, February 28). Graduate admission variables and future success. *Science, 187,* 721–723.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on a myth.* New York: Free Press.

Dawis, R. V. (1992). The individual differences tradition in counseling psychology. *Journal of Counseling Psychology, 39,* 7–19.

Dawis, R. V., & Lofquist, L. H. (1984). *A psychological theory of work adjustment.* Minneapolis: University of Minnesota Press.

Dunnette, M. D. (1966). Fads, fashions, and folderol in psychology. *American Psychologist, 21,* 343–352.

Ehrlich, S. B., Levine, S. C., & Goldin-Meadow, S. (2006). The importance of gesture in children's spatial reasoning. *Developmental Psychology, 42,* 1259–1268.

Eliot, J. C. (1987). Models of psychological space: Psychometric, developmental, and experimental approaches. New York: Springer-Verlag.

Eliot, J. C., & Smith, I. M. (1983). *An international dictionary of spatial tests.* Windsor, England: NFER-Nelson.

Flanagan, J. C., Dailey, J. T., Shaycoft, M. F., Gorman, W. A., Orr, D. B., & Goldberg, I. (1962). *Design for a study of American youth.* Boston: Houghton Mifflin.

Friedman, T. L. (2005). *The world is flat: A brief history of the twenty-first century.* New York: Farrar, Straus & Giroux.

Geary, D. C. (1998). *Male, female: The evolution of human sex differences.* Washington, DC: American Psychological Association.

Geary, D. C. (2005). *The origin of mind: Evolution of brain, cognition, and general intelligence.* Washington, DC: American Psychological Association.

Gohm, C. L., Humphreys, L. G., & Yao, G. (1998). Underachievement among spatially gifted students. *American Educational Research Journal, 35,* 515–531.

Gottfredson, L. S. (2003). The challenge and promise of cognitive career assessment. *Journal of Career Assessment, 11,* 115–135.

Gottfredson, L. S. (2005). Using Gottfredson's theory of circumscription and compromise in career guidance and counseling. In S. D. Brown & R. W. Lent (Eds.), *Career development and counseling: Putting theory and research to work* (pp. 71–100). New York: Wiley.

Hackman, D. A., & Farah, M. J. (2009). Socioeconomic status and the developing brain. *Trends in Cognitive Sciences, 13,* 65–73.

Hegarty, M., & Waller, D. A. (2005). Individual differences in spatial abilities. In P. Shah & A. Miyake (Eds.), *The Cambridge handbook of visuospatial thinking* (pp. 121–169). New York: Cambridge University Press.

Humphreys, L. G. (1991). Some unconventional analyses of resemblance coefficients for male and female monozygotic and dizygotic twins. In D. Cicchette & W. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul Everett Meehl* (pp. 158–187). Minneapolis: University of Minnesota Press.

Humphreys, L. G., & Lubinski, D. (1996). Brief history and psychological significance of assessing spatial visualization. In C. P. Benbow & D. Lubinski (Eds.), *Intellectual talent: Psychometric and social issues* (pp. 116–140). Baltimore: Johns Hopkins University Press.

Humphreys, L. G., Lubinski, D., & Yao, G. (1993). Utility of predicting group membership and the role of spatial visualization in becoming an engineer, physical scientist, or artist. *Journal of Applied Psychology, 78,* 250–261.

Johnson, W., & Bouchard, T. J. (2007a). Sex differences in mental abilities: *g* masks the dimensions on which they lie. *Intelligence, 35,* 23–39.

Johnson, W., & Bouchard, T. J. (2007b). Sex differences in mental ability: A proposed means to link them to brain structure and function. *Intelligence, 35,* 197–209.

Keating, D. P., & Stanley, J. S. (1972). Extreme measures for the mathematically gifted in mathematics and science. *Educational Researcher, 1,* 3–7.

Levine, S. C., Vasilyeva, M., Lourenco, S. F., Newcombe, N. S., & Huttenlocher, J. (2005). Socioeconomic status modifies the sex difference in spatial skill. *Psychological Science, 16,* 841–845.

Lohman, D. F. (1988). Spatial abilities as traits, processes, and knowledge. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 4, pp. 181–248). Hillsdale, NJ: Erlbaum.

Lohman, D. F. (1994a). Spatial ability. In R. J. Sternberg (Ed.), *Encyclopedia of intelligence* (Vol. 2, pp. 1000–1007). New York: Macmillan.

Lohman, D. F. (1994b). Spatially gifted, verbally inconvenienced. In N. Colangelo, S. G. Assouline, & D. L. Ambroson (Eds.), *Talent development: Vol. 2. Proceedings from the 1993 Henry B. and Jocelyn Wallace National Research Symposium on Talent Development* (pp. 251–264). Dayton, OH: Ohio Psychology Press.

Lohman, D. F. (1996). Spatial ability and G. In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and assessment* (pp. 97–116). Hillsdale, NJ: Erlbaum.

Lohman, D. F. (2005). The role of nonverbal ability tests in identifying academically gifted students: An aptitude perspective. *Gifted Child Quarterly, 49,* 111–138.

Lubinski, D. (2000). Scientific and social significance of assessing individual differences: "Sinking shafts at a few critical points." *Annual Review of Psychology, 51,* 405–444.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904). "'General intelligence,' objectively determined and measured." *Journal of Personality and Social Psychology, 86,* 96–111.

Lubinski, D., & Benbow, C. P. (2000). States of excellence. *American Psychologist, 55,* 137–150.

Lubinski, D., & Benbow, C. P. (2006). Study of Mathematically Precocious Youth after 35 years: Uncovering antecedents for the development of math–science expertise. *Perspectives on Psychological Science, 1,* 316–345.

Lubinski, D., & Dawis, R. V. (1992). Aptitudes, skills, and proficiencies. In M. D. Dunnette & L. M. Hough (Eds.), *The handbook of industrial/organizational psychology* (2nd ed., pp. 1–59). Palo Alto, CA: Consulting Psychologists Press.

Lubinski, D., & Humphreys, L. G. (1990a). A broadly based analysis of mathematical giftedness. *Intelligence, 14,* 327–355.

Lubinski, D., & Humphreys, L. G. (1990b). Assessing spurious "moderator effects": Illustrated substantively with the hypothesized ("synergistic") relation between spatial visualization and mathematical ability. *Psychological Bulletin, 107,* 385–393.

Lubinski, D., & Humphreys, L. G. (1996). Seeing the forest from the trees: When predicting the behavior or status of groups, correlate means. *Psychology, Public Policy, and Law, 2,* 363–376.

Lubinski, D., & Humphreys, L. G. (1997). Incorporating general intelligence into epidemiology and the social sciences. *Intelligence, 24,* 159–201.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Ciccetti & W. Grove (Eds.), *Thinking clearly about psychology* (pp. 3–39). Minneapolis: University of Minnesota Press.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141.

National Academy of Sciences. (2005). *Rising above the gathering storm.* Washington, DC: National Academy Press.

Newcombe, N. S., & Uttal, D. H. (2006). Whorf versus Socrates, round 10. *Trends in Cognitive Sciences, 10,* 394–396.

Noble, K. G., McCandliss, B. D., & Farah, M. J. (2007). Socioeconomic gradients predict individual differences in neurocognitive abilities. *Developmental Science, 10,* 464–480.

Park, G., Lubinski, D., & Benbow, C. P. (2007). Contrasting intellectual patterns for creativity in the arts and sciences: Tracking intellectually precocious youth over 25 years. *Psychological Science, 18,* 948–952.

Park, G., Lubinski, D., & Benbow, C. P. (2008). Ability differences among people who have commensurate degrees matter for scientific creativity. *Psychological Science, 19,* 957–961.

Paterson, D. G. (1957). The conservation of human talent. *American Psychologist, 12,* 134–144.

Paterson, D. G., Elliott, R. M., Anderson, L. D., & Toops, H. A. (1930). *Minnesota Mechanical Ability Tests.* Minneapolis: University of Minnesota Press.

Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research, 37,* 239–253.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85,* 112–118.

Schmidt, D. B., Lubinski, D., & Benbow, C. P. (1998). Validity of assessing educational–vocational preference dimensions among intellectually talented 13-year-olds. *Journal of Counseling Psychology, 45,* 436–453.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Shaycoft, M. F. (1967). *The high school years: Growth in cognitive skills.* Pittsburgh, PA: School of Education, University of Pittsburgh, American Institutes for Research.

Shea, D. L., Lubinski, D., & Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *Journal of Educational Psychology, 93,* 604–614.

Silverman, I., Choi, J., & Peters, M. (2007). The hunter–gatherer theory of sex differences in spatial abilities: Data from 40 countries. *Archives of Sexual Behavior, 36,* 261–268.

Smith, I. M. (1964). *Spatial ability: Its educational and social significance.* London: University of London Press.

Snow, R. E. (1999). Commentary: Expanding the breadth and depth of admissions testing. In S. Messick (Ed.), *Assessment in higher education* (pp. 133–140). Hillsdale, NJ: Erlbaum.

Snow, R. E., Corno, L., & Jackson, D. N., III. (1996). Individual differences in affective and conative functions. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 243–310). New York: MacMillan.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–332). New York: Macmillan.

Stanley, J. C. (1996). In the beginning: The Study of Mathematically Precocious Youth. In C. P. Benbow & D. Lubinski (Eds.), *Intellectual talent* (pp. 225–235). Baltimore: Johns Hopkins University Press.

Stanley, J. C. (2000). Helping students learn only what they don't already know. *Psychology, Public Policy, and Law, 6,* 216–222.

Steen, L. A. (1988, April 29). The science of patterns. *Science, 240,* 611–616.

Super, D. E., & Bachrach, P. B. (1957). *Scientific careers and vocational development theory.* New York: Bureau of Publications, Teachers College, Columbia University.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills, 47,* 599–604.

Vernon, P. A. (Ed.). (1961). *The structure of human abilities* (2nd ed.). London: Methuen London.

Wai, J., Lubinski, D., & Benbow, C. P. (2005). Creativity and occupational accomplishments among intellectually precocious youth: An age 13 to age 33 longitudinal study. *Journal of Educational Psychology, 94,* 785–794.

Webb, R. M., Lubinski, D., & Benbow, C. P. (2007). Spatial ability: A neglected dimension in talent searches for intellectually precocious youth. *Journal of Educational Psychology, 99,* 397–420.

Wise, L. L., McLaughlin, D. H., & Steel, L. (1979). *The Project TALENT data bank.* Palo Alto, CA: American Institutes for Research.

(Appendixes follow)

Appendix A

Breakdown by Sex of the Degrees and Occupations Included in Figure 5 and Appendix B

Degrees

Engineering
- Engineering (1137/336/71, 6/3/0)

Physical science
- Biochemistry (8/9/21, 8/8/5)
- Chemistry (244/43/100, 95/10/9)
- Physical science (122/37/9, 21/5/0)
- Physics (131/69/58, 4/1/0)

Math/computer science
- Computer science (27/42/6, 2/9/0)
- Math (474/128/42, 365/74/3)
- Statistics (7/11/6, 2/2/0)

Biological science
- Anatomy (12/11/16, 8/6/3)
- Biological science (332/69/28, 209/39/6)
- Botany (15/15/7, 7/11/3)
- Zoology (112/20/14, 45/11/2)

Humanities
- English (344/110/24, 1056/140/6)
- Foreign language (111/37/7, 330/84/3)
- History (635/146/24, 374/62/3)
- Humanities (12/3/0, 21/5/1)
- Journalism (89/34/0, 53/8/0)
- Philosophy (110/21/8, 29/1/2)
- Religion (30/39/4, 32/5/0)

Social science
- Economics (410/62/16, 51/3/2)
- Political science (359/72/17, 147/19/3)
- Psychology (338/142/71, 311/66/26)
- Social science (176/43/5, 195/21/4)
- Sociology (234/38/12, 388/18/2)

Visual arts
- Architecture (53/6/0, 6/0/0)
- Fine arts (88/29/0, 219/29/0)

Occupations

Engineering
- Engineer (NEC) (219, 1)
- Civil engineer (163, 2)
- Electrical engineer (315, 0)
- Mechanical engineer (192, 1)
- Aeronautical engineer (60, 0)
- Chemical engineer (78, 1)

Physical science
- Physical scientist (NEC) (20, 1)
- Chemist (153, 19)
- Physicist (62, 2)
- Astronomer (1, 0)
- Geologist (38, 3)
- Meteorologist (10, 0)
- Biochemist (26, 12)

Math/computer science
- Mathematician (38, 10)
- Statistician (28, 9)
- Systems analyst (258, 68)
- Computer programmer (175, 61)
- Computer specialist (NEC) (63, 1)

Biological science
- Biological scientist (NEC) (23, 10)
- Pharmacologist (8, 1)
- Microbiologist (19, 16)

Humanities
- Writer (NEC) (15, 8)
- Fiction writer (1, 1)
- Nonfiction writer (16, 20)
- Journalists/reporters (31, 16)
- Radio–TV reporter (6, 6)
- Publisher (4, 2)
- Editor (39, 39)

Visual arts
- Architect (44, 1)
- Painter/sculptor (4, 5)
- Commercial/fashion artist (18, 24)
- Draftsman (13, 1)
- Photographer (15, 3)
- Interior designer (3, 11)
- Landscape architect (10, 0)

Performing arts
- Musician (instrumental) (15, 9)
- Singer (4, 0)
- Music-related (NEC) (5, 6)
- Dancer, choreographer (0, 9)
- Actor (1, 4)
- Theatrical director (7, 1)
- Theater occupation (NEC) (20, 7)
- Performer (NEC) (2, 0)
- Radio or TV announcer (3, 0)
- Performing artist (NEC) (6, 5)

Business
- Executive (NEC) (13, 1)
- In business for self (NEC) (37, 2)
- Industry, business, or commerce (3, 0)
- Real estate (NEC) (9, 0)
- Insurance (NEC) (3, 0)
- Market analyst (29, 8)
- Banking and finance (196, 8)
- Investment consultant (39, 5)
- CPA (282, 7)
- Accountant or auditor (406, 22)
- Purchasing and procurement (96, 8)
- Buyer for retail store (23, 9)
- Efficiency expert (NEC) (242, 3)
- Contractor (62, 0)
- Franchiser (123, 26)
- Salesclerk/cashier (42, 31)
- Routeman (5, 0)
- Stockbroker (53, 3)
- Securities dealer (15, 0)
- Real estate salesman (77, 15)
- Insurance salesman (137, 1)
- Auto salesman (10, 0)
- Other salesmen (NEC) (560, 24)
- Sales manager (241, 15)

Education
- Teaching (NEC) (269, 352)
- Teaching young children (1, 9)
- Teaching preschool children (2, 165)
- Teaching elementary school (273, 1232)
- Teaching high school (NEC) (173, 173)
- Teaching high school math (195, 103)
- Teaching high school science (236, 74)
- Teaching high school social studies (205, 96)
- Teaching high school English (122, 259)
- Teaching high school foreign language (51, 76)
- Teaching high school commercial education (49, 63)
- Teaching high school home economics (0, 84)
- Teaching high school trade education (141, 11)
- Teaching high school physical education (117, 101)
- Teaching art (46, 67)
- Teaching music (55, 78)
- Teaching speech in high school (8, 9)
- Teaching the handicapped (71, 147)
- Speech therapist (11, 38)
- School administrator (not college) (124, 40)
- Reading specialist (16, 54)

Appendix A (*continued*)

Degrees	Occupations		
Performing arts	Translator/linguist (4, 4)	Advertiser (47, 17)	Other education specialist (46, 39)
Music (63/40/3, 128/28/1)	University teacher: English (53, 37)	Public relations (41, 34)	Teacher's aide (13, 44)
Performing arts (18/18/1, 40/16/0)	University teacher: foreign language (16, 19)	Personnel administrator (233, 53)	Medicine
Business	Social science	Appraiser/estimator (118, 15)	MD general practitioner (34, 0)
Accounting (708/32/4, 36/1/1)	Psychologist (97, 39)	Credit investigator (137, 15)	MD surgeon (56, 2)
Business and commerce (1393/134/9, 249/9/1)	Economist (17, 3)	Manager and administrator (NEC) (515, 50)	MD psychiatrist (38, 11)
Education	Sociologist (1, 0)	Manufacturing manager (56, 1)	MD medical researcher (12, 6)
Education other (232/441/21, 378/425/2)	Social scientist (NEC) (40, 22)	Retail trade manager (169, 18)	MD other and unspecified (338, 21)
Elementary education (164/80/1, 2149/414/0)	University teacher: social science (109, 29)	Private business agent (3, 0)	Dentistry
Physical education (258/66/2, 222/51/2)		Developer (14, 0)	Dentist (189, 1)
Specific doctorates		Business supervisor (117, 6)	Law
JD (939, 42); DDS (214, 1); MD (490, 48); EdD (43, 17)			Lawyer (777, 28)

Note. This table includes the breakdown by sex of the degrees and occupations included in Figure 5 and Appendix A. In the degrees column, the respective sample sizes are given for bachelors, masters, and doctorates for males and females, respectively (B/M/D for males, B/M/D for females). In the occupations columns, the sample sizes (males/females) are reported. The specific doctorates category in the degrees column and the occupations columns pertain only to Figure 5. Figure B1 includes data from the remainder of the degrees column (i.e., Engineering through Education). NEC = not elsewhere classified.

Appendix B

Average *Z* Scores of Participants on Both General Ability Level and Spatial, Mathematical, and Verbal Ability Level for Bachelor's Degrees, Master's Degrees, and PhD Degrees Plotted by Field

It is important to note the importance of spatial ability for those securing degrees in math/computer science, physical science, and engineering. Hegarty and Waller (2005, p. 155) discussed the importance of spatial ability in the performance of surgeons. Bingham (1937) anticipated this topic, noting that, for surgeons and dentists,

> quite as indispensable is aptitude for visualizing vividly in three dimensions; for it is necessary to see in their true positions and to manipulate the forms observed in a dentist's little mirror or in a laryngoscope; also to picture correctly the highly complicated unseen structures beneath the body surface—arteries, nerves, muscles, tendons, joints, glands, vital organs—perhaps at the end of a probe. (p. 172)

We conducted an analysis using the surgeons and other MDs that can be found here and in Appendix A (MD surgeon, $n = 58$; MD all others, $n = 460$). The difference between the surgeons (avg. $z = 1.17$) and the remainder (avg. $z = 1.12$) on spatial ability was 0.05. The highest in spatial ability were the MD medical researchers (avg. $z = 1.27$) in comparison to all other subgroups. Spatial ability is evidently important not only for surgeons but all the medical fields examined in Project TALENT, and in particular for medical research.

We conducted an analysis to determine the similarity between all four cohorts compared to the 9th-grade cohort alone. The average difference across all four grades combined and the 9th-grade sample was less than the absolute value of 0.08. For completeness, the *g* level (average of S + M + V) of the bachelors (BA and BS) and doctorates (PhDs) was computed within each group. In the order corresponding to the graph, these were as follows: engineering (PhD = 1.73; bachelors = 1.22), physical science (PhD = 1.62; bachelors = 1.15), math/computer science (PhD = 1.75; bachelors = 1.18), biological science (PhD = 1.33; bachelors = 0.86), humanities (PhD = 1.34; bachelors = 0.84), social science (PhD = 1.29; bachelors = 0.75), arts (masters + PhD = 0.97; bachelors = 0.71), business (masters + PhD = 0.99; bachelors = 0.64), and education (masters + PhD = 0.64; bachelors = 0.46).

(*Appendixes continue*)

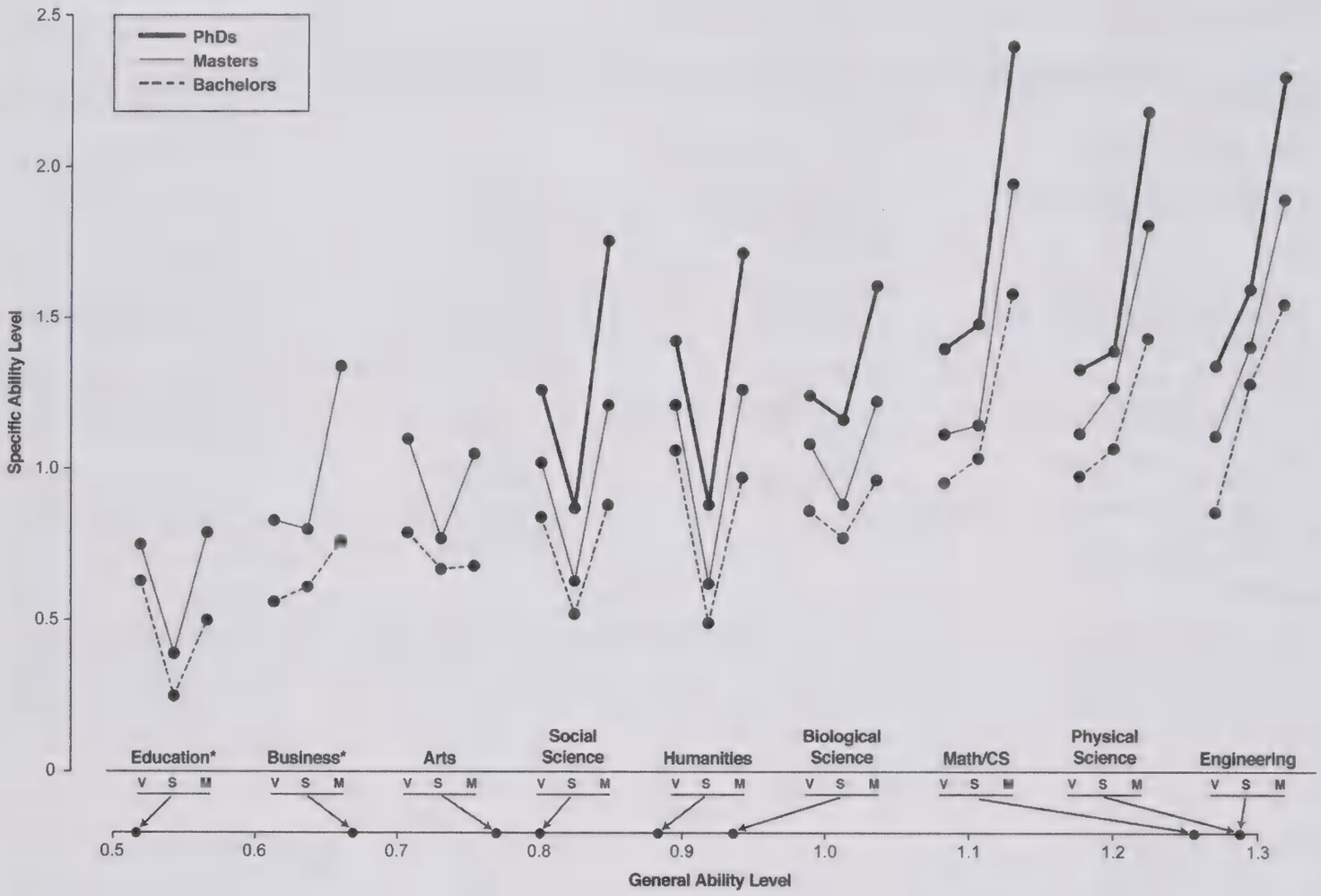


Figure B1. * For education and business, masters and doctorates were combined because the doctorate samples for these groups were too small to obtain stability ($n < 30$). For the specific n for each degree by sex that composed the major groupings, see Appendix A. Average z scores of participants on spatial, mathematical, and verbal ability for bachelor's degrees, master's degrees, and PhDs are plotted by field in Figure B1. The groups are plotted in rank order of their normative standing on g (verbal [V] + spatial [S] + mathematical [M]) along the x-axis, and each arrow indicates on the continuous scale where each field lies on general mental ability. All x-axis values are based on the weighted means across each degree grouping. This figure is standardized in relation to all participants with complete ability data at the time of initial testing. Respective ns for each group (males + females) were as follows (for bachelor's, master's, and doctorates, respectively): engineering (1,143, 339, 71), physical science (633, 182, 202), math/computer science (877, 266, 57), biological science (740, 182, 79), humanities (3,226, 695, 82), social science (2,609, 484, 158), arts (615, masters + doctorates = 171), business (2,386, masters + doctorates = 191), and education (3,403, masters + doctorates = 1,505).

Appendix C

Spatial Ability Composite Means and Standard Deviations by Socioeconomic Status (SES) Quartile, Grade, and Sex

SES quartile	9th grade		10th grade		11th grade		12th grade	
	Males	Females	Males	Females	Males	Females	Males	Females
1	60.67 (20.70)	50.99 (17.66)	65.16 (21.68)	54.36 (18.98)	69.95 (22.16)	57.57 (19.52)	72.79 (22.73)	60.05 (20.10)
	13,056	12,196	12,642	12,784	11,502	12,101	9,263	9,900
2	68.06 (21.10)	58.68 (18.54)	73.77 (21.19)	62.57 (19.66)	78.68 (21.63)	65.22 (19.31)	81.90 (21.55)	67.75 (19.83)
	11,514	11,509	11,890	11,825	11,095	10,167	9,504	10,609
3	72.75 (21.31)	63.43 (18.77)	78.33 (21.11)	67.21 (19.38)	83.03 (20.86)	69.55 (19.54)	86.30 (21.02)	71.82 (19.73)
	11,512	11,781	11,477	10,046	9,187	11,581	9,493	8,705
4	78.80 (20.92)	69.23 (18.98)	83.97 (20.69)	72.20 (19.77)	87.59 (20.61)	74.98 (19.78)	91.52 (20.33)	76.57 (20.12)
	11,123	11,883	10,696	11,147	10,368	10,256	8,430	9,432

Note. In each cell, the mean, standard deviation (in parentheses) and *n* are reported.

Received December 20, 2007
Revision received September 26, 2008
Accepted April 10, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 836–852

0022-0663/09/$12.00 DOI: 10.1037/a0016026

The Importance of Prior Knowledge When Comparing Examples: Influences on Conceptual and Procedural Knowledge of Equation Solving

Bethany Rittle-Johnson
Vanderbilt University

Jon R. Star
Harvard University

Kelley Durkin
Vanderbilt University

Comparing multiple examples typically supports learning and transfer in laboratory studies and is considered a key feature of high-quality mathematics instruction. This experimental study investigated the importance of prior knowledge in learning from comparison. Seventh- and 8th-grade students (N = 236) learned to solve equations by comparing different solution methods to the same problem, comparing different problem types solved with the same solution method, or studying the examples sequentially. Unlike in past studies, many students did not begin the study with equation-solving skills, and prior knowledge of algebraic methods was an important predictor of learning. Students who did not attempt algebraic methods at pretest benefited most from studying examples sequentially or comparing problem types, rather than from comparing solution methods. Students who attempted algebraic methods at pretest learned more from comparing solution methods. Students may need sufficient prior knowledge in a domain before they benefit from comparing alternative solution methods. These findings are in line with findings on the expertise-reversal effect.

Keywords: comparison, conceptual (declarative) knowledge, procedural knowledge, flexibility, mathematics education

Comparing multiple examples typically supports learning and transfer in laboratory studies and is considered a key feature of high-quality mathematics instruction (e.g., Ball, 1993; Gentner, Loewenstein, & Thompson, 2003; Gick & Holyoak, 1983; Silver, Ghousseini, Gosen, Charalambous, & Strawhun, 2005). Indeed, according to Gentner (2005), "The simple, ubiquitous act of comparing two things is often highly informative to human learners .... Comparison is a general learning process that can promote deep relational learning and the development of theory-level explanations" (pp. 247, 251). However, little is known about the constraints on when comparison is most effective. For example, is comparison effective early in the learning process or only after learners have sufficient prior knowledge in the domain? In the current study, we evaluated whether prior knowledge of solution methods impacted the effectiveness of comparison for supporting mathematics learning.

Experimental Research on Comparison

Experimental studies on comparison have shown that comparing two examples supports greater learning than does studying a single example or studying two examples independently. This has been shown across a variety of tasks, ranging from infants learning perceptual categories (Oakes & Ribar, 2005) to preschoolers learning words (Gentner & Namy, 2004) to master's-level students learning contract negotiation strategies (Gentner et al., 2003). In this last instance, students who were prompted to compare two scenarios in which the same negotiation strategy was illustrated were much more likely to transfer the solution strategy to a new case than were students who read and reflected on the examples independently (Gentner et al., 2003). Across studies, learners who compare multiple examples are better able to generalize the information to new instances (e.g., Catrambone & Holyoak, 1989; Cummins, 1992; Gentner, Loewenstein, & Hung, 2007; Gick & Holyoak, 1983; VanderStoep & Seifert, 1993).

However, little empirical research has evaluated the benefits of comparison in school-age children or with academic tasks. Recent work has attempted to address both of these limitations. For example, students who compared examples of two methods for

Bethany Rittle-Johnson and Kelley Durkin, Department of Psychology and Human Development, Peabody College, Vanderbilt University; Jon R. Star, Graduate School of Education, Harvard University.

This research was supported with funding from the Institute of Education Sciences, U.S. Department of Education, Grants R305H050179 and R305B040110. The opinions expressed are those of the authors and do not represent the views of the U.S. Department of Education. A special thanks to the students and teachers at Forest Grove Middle School for participating in this research. Thanks to Neil Heffernan, Cristina Heffernan, Christine Yang, Martina Olzog, Jennifer Rabb, Theodora Chang, Nira Gautam, Natasha Perova, Leena Razzaq, Mingyu Feng, Holly Harris, Shanelle Chambers, Anna Krueger, Heena Ali, Kelly Cashen, Erin Fahey, Adam Porter, and Calie Traver for help in collecting and coding the data. Thanks to E. Warren Lambert of the Vanderbilt Kennedy Center for statistical help.

Correspondence concerning this article should be addressed to Bethany Rittle-Johnson, 230 Appleton Place, Peabody No. 0552, Nashville, TN 37203. E-mail: b.rittle-johnson@vanderbilt.edu

solving equations such as $3(x + 5) = 15$ with a partner during math class were better able to solve equations and to flexibly use and evaluate nonstandard solution methods than were students who studied the same examples one at a time (Rittle-Johnson & Star, 2007). These findings have been replicated with fifth graders learning about computational estimation (Star & Rittle-Johnson, 2009).

Comparison in Mathematics Classrooms

Despite the paucity of experimental research on comparison with school-age children or with academic tasks, mathematics educators do incorporate comparison in their instruction. Expert mathematics teachers in the United States have students share and compare solution methods (e.g., Ball, 1993; Lampert, 1990), as do teachers in high-performing countries, such as Japan (Richland, Holyoak, & Stigler, 2004; Richland, Zur, & Holyoak, 2007). As a result, this instructional practice is emphasized in the National Council of Teachers of Mathematics (NCTM) *Standards* (1989, 2000).

Expert teachers are not the only ones who use comparison. In a representative sample, teachers from the United States, Japan, and Hong Kong made comparisons multiple times in their lessons (Richland et al., 2004; Richland et al., 2007). However, U.S. teachers frequently used comparison in ways that did not seem conducive to student learning. For example, they sometimes made analogies between examples that shared limited similarities, compared examples that were not visually available, or did not provide visual or gestural cues to help map between the two examples when both were visually available.

Role of Prior Knowledge in Learning From Comparison

Although comparing examples can improve learning, it is unlikely that comparison is always an optimal instructional approach. Classic research on aptitude–treatment interactions (Snow, 1992) and more recent research on expertise-reversal effects (see Kalyuga, 2007, for a review) indicate that characteristics of the learner influence which instructional approaches are most effective. The literature on the expertise-reversal effect focuses on learners' levels of prior knowledge in particular; a reversal effect occurs when level of prior knowledge interacts with the effectiveness of different instructional techniques (Kalyuga, 2007). For example, studying worked examples is more effective than solving problems for novices in a domain, but the reverse is true for learners with a moderate amount of problem-solving knowledge in the domain (Kalyuga & Sweller, 2004). Level of expertise has also been shown to interact with representational format of the instructional materials (e.g., verbal, pictorial, or computer simulations), amount and type of direct instruction, and type of practice (Kalyuga, 2007; Rittle-Johnson & Kmicikewycz, 2008). For novices, tasks can easily overload their working memory, as they must deal with many new elements of information at once. In contrast, learners with some experience in a domain can use their existing knowledge structures to interpret and complete the task without overloading their working memory. This helps to explain why constrained tasks and high levels of instructional guidance are often needed to facilitate learning in novices, whereas more experienced learners often benefit from more complex and less guided tasks (Cronbach & Snow, 1977; Kalyuga, 2007).

Although no prior studies have explicitly examined this issue, it seems plausible that prior knowledge also plays an important role in the effectiveness of comparison. For novices in a domain, comparison of worked examples requires interpreting each example as well as comparing the similarities and differences between the examples. Therefore, we predicted that comparing two examples would not be very effective for learners with little prior knowledge in a domain, even though it has been shown to be effective at supporting learning for those with moderate prior knowledge (Rittle-Johnson & Star, 2007, in press; Star & Rittle-Johnson, 2009). Exploratory findings from one of our previous studies suggested that students with low prior knowledge did not benefit from comparing solution methods (Albro et al., 2007). In addition, one study found that students with low levels of procedural knowledge learned more from sequential presentation of information than from concurrent presentation of the information, whereas students with moderate levels of procedural knowledge learned equally well from the two presentation formats (Clarke, Ayres, & Sweller, 2005). Students were not making comparisons in the concurrent condition, but these findings nevertheless suggest that sequential presentation of examples might be preferable for low-knowledge learners. Additional support for the hypothesis that novices may benefit less from comparing examples than experts do comes from research on analogical reasoning. Learning from comparing unfamiliar examples is often difficult for young children (Gentner et al., 2007; Kotovsky & Gentner, 1996) and for college students who do not receive additional instructional support (Schwartz & Bransford, 1998).

Target Domain and Outcomes

We investigated the importance of prior knowledge for learning from comparison with middle-school students learning to solve equations. Many people in mathematics education consider linear equation solving a "basic skill" (National Mathematics Advisory Panel, 2008). In fact, the NCTM recommends linear equation solving as a curriculum focal point for Grade 7 (NCTM, 2006). Regrettably, students often memorize rules and do not learn flexible and meaningful ways to solve equations (Kieran, 1992).

Students need opportunities to learn multiple ways to solve equations and how to choose flexibly among the methods (Beishuizen, van Putten, & van Mulken, 1997; Blöte, Van der Burg, & Klein, 2001; Star & Seifert, 2006). We focused on multiple-step linear equations, such as $3(x + 1) = 15$, that can be solved in multiple ways. We assessed students' competence with equations for three critical components of mathematics knowledge: procedural knowledge, procedural flexibility, and conceptual knowledge (Kilpatrick, Swafford, & Findell, 2001). Procedural knowledge can be defined as the ability to execute action sequences to solve problems (Hiebert & Wearne, 1996; Rittle-Johnson, Siegler, & Alibali, 2001). For our procedural knowledge measure, we included problems with familiar as well as unfamiliar problem features in order to assess learning of correct procedures and ability to adapt the procedures to new problem features (Paas & Van Merrienboer, 1994; Singley & Anderson, 1989). Procedural flexibility can be defined as knowing how to solve a problem in

multiple ways and when each way is most efficient (Kilpatrick et al., 2001; Star, 2005). Our assessment included both an independent measure of flexibility knowledge and a measure of flexible use of solution methods on the procedural knowledge problems (Star & Rittle-Johnson, 2008). This was done to differentiate between knowledge and use of flexible procedures, as children sometimes know more sophisticated procedures but do not implement them (Bjorklund, Miller, Coyle, & Slawinski, 1997). Finally, conceptual knowledge can be defined as "an integrated and functional grasp of mathematical ideas" (Kilpatrick et al., 2001, p. 118). Our conceptual knowledge measure focused on students' abilities to recognize and explain key domain concepts (Carpenter, Franke, Jacobs, Fennema, & Empson, 1998; Hiebert & Wearne, 1996).

Current Study: Importance of Prior Knowledge for Learning From Comparison

We evaluated the effects of comparison relative to sequential study of the same examples with middle-school students who varied in their prior knowledge of equation solving. We included two different comparison conditions that reflected the two primary forms of comparison in the literature. The first was comparing the same problem solved with two different algebraic solution methods (*compare methods*; see Figure 1). This form of comparison is recommended in mathematics education standards (NCTM, 2000) and has been shown to improve mathematics learning (Rittle-Johnson & Star, 2007, in press; Star & Rittle-Johnson, 2009). The second was comparing different problem types solved with the same solution method (*compare problems*; see Figure 1). This form of comparison is commonly used in cognitive science research; learners typically compare examples with different surface features but the same underlying solution method (e.g., Catrambone & Holyoak, 1989; Cummins, 1992; Gentner et al., 2003; Gick & Holyoak, 1983; VanderStoep & Seifert, 1993). Comparing problem types has been shown to support some components of mathematics learning (Rittle-Johnson & Star, in press). In the sequential condition, students studied the same examples as those in the other two conditions, but the examples were presented one at a time on separate pages (see Figure 1). The students in the sequential condition were not prompted to make comparisons between examples.

Students in this study were expected to vary in their prior knowledge of equation-solving procedures. In our past studies, participating students were in prealgebra courses, so they had learned and practiced related procedures in prior lessons and often spontaneously used one of our target solution methods at pretest (Rittle-Johnson & Star, 2007, in press). In this study, we worked with students using a general middle-school mathematics curriculum. In particular, the students were using a reform-oriented, concept-based curriculum that focused less on gradually building skills and more on relevant concepts. Most students had completed a unit on understanding simple linear equations, but they had limited experience with equation solving. Some teachers had chosen to expose some, but not all, of their students to multistep equations, so students were expected to vary in their knowledge of algebraic solution methods for multistep equations, but few were expected to be proficient in equation solving.

We hypothesized, on the basis of the expertise-reversal effect, that the effect of condition would interact with students' prior knowledge of algebraic solution methods (Kalyuga, 2007). In particular, we predicted that (a) students with little prior knowledge of algebraic procedures would learn the most from sequential study of examples and learn less from either type of comparison, because sequential study is thought to impose less working memory load and has been shown to benefit learners with low prior knowledge (Clarke et al., 2005); and (b), based on our previous findings with more experienced problem solvers, students with some knowledge of algebraic procedures would learn the most from comparing solution methods and learn the least from sequential study of examples (Rittle-Johnson & Star, 2007, in press).

Method

Participants

Participants were drawn from 11 classrooms at a low-performing, urban middle school in Massachusetts. All four eighth-grade math teachers and one seventh-grade math teacher at the school volunteered to participate, and the teachers identified classes they felt were prepared to learn about multistep equations. Nine of the classes were eighth-grade classes (5 regular and 4 honors-level classes) and 2 of the classes were seventh-grade classes (both honors level). Students were using the Connected Mathematics 2 curriculum (Lappan, Fey, Fitzgerald, Friel, & Phillips, 2009), and the eighth graders had completed the Moving Straight Ahead unit on linear equations before participating. Most teachers reported spending only a few days on linear equation solving. In 5 of the 11 classes, teachers reported introducing some of their students to multistep equations and distribution across parentheses, even though it was not covered in the curriculum unit, but indicated that these students had had little opportunity to practice solving multistep equations.

All 239 students (136 female) from these classes participated in the fall. Of these students, 3 were dropped from the analysis, 1 because she was absent for two of the three intervention sessions and 2 because they refused to complete the intervention packets (they were partners). The remaining 236 students were 45 seventh-grade students and 191 eighth-grade students, with 121 of the eighth-grade students in the honors-level classes. The average age was 13.3 years (range 11.9–15.7 years), and a majority were Caucasian (72%), with equal numbers of Hispanic, African American, and Asian students (9% of students were of each race/ethnicity). The school used the Measures of Academic Progress as a norm-referenced test to measure mathematics achievement and growth. The average score for participating students was in the 74th percentile, but there was great variability, with scores ranging from the 12th to the 99th percentile.

Design

We employed a pretest–intervention–posttest design with a retention test. Within each classroom, pairs of students were randomly assigned to compare solution methods (abbreviated as compare methods; $n = 80$), to compare problem types (abbreviated as compare problems; $n = 78$), or to study examples sequentially, without comparison (abbreviated as *sequential*; $n = 78$). During the 3-day intervention, students studied the worked examples with a partner and answered explanation prompts designed to guide attention to the example features targeted in each condition. We chose to have stu-

A. Compare Solution Methods

Nathan's Solution:		**Patrick's Solution:**	
$5(y + 1) = 3(y + 1) + 8$		$5(y + 1) = 3(y + 1) + 8$	
$5y + 5 = 3y + 3 + 8$	*Distribute* _____	$2(y + 1) = 8$	*Subtract* $3(y + 1)$ *on Both*
$5y + 5 = 3y + 11$	*Combine* _____	$y + 1 = 4$	*Divide* _____ *on Both*
$2y + 5 = 11$	*Subtract* _____ *on Both*	$y = 3$	*Subtract* _____ *on Both*
$2y = 6$	*Subtract* _____ *on Both*		
$y = 3$	*Divide* _____ *on Both*		

5. Describe 2 ways that these students' solutions are **different.**

6. What must be true about an equation for Patrick's way to be easier than Nathan's way?

_____

B. Compare Problem Types

Abby's Solution:		**Patrick's Solution:**	
$3(h - 2) + 5(h - 2) = 24$		$5(y + 1) = 3(y + 1) + 8$	
$8(h - 2) = 24$	*Combine* _____	$2(y + 1) = 8$	*Subtract* $3(y + 1)$ *on Both*
$h - 2 = 3$	*Divide* _____ *on Both*	$y + 1 = 4$	*Divide* _____ *on Both*
$h = 5$	*Add* _____ *on Both*	$y = 3$	*Subtract* _____ *on Both*

5. Describe one way the students' **problems** are the same and one way they are different.

6. Patrick's first step is **different** from Abby's first step because:

_____

C. No Compare (Sequential)

Abby's Solution:	
$3(h - 2) + 5(h - 2) = 24$	
$8(h - 2) = 24$	*Combine* _____
$h - 2 = 3$	*Divide* _____ *on Both*
$h = 5$	*Add* _____ *on Both*

Why did Abby combine like terms for her first step?

------NEXT PAGE------

Patrick's Solution:	
$5(y + 1) = 3(y + 1) + 8$	
$2(y + 1) = 8$	*Subtract* _____ *on Both*
$y + 1 = 4$	*Divide* _____ *on Both*
$y = 3$	*Subtract* _____ *on Both*

Could you use Patrick's way to solve many different kinds of problems? **Circle one:** YES NO Why or why not?

Figure 1. Sample page of the intervention packet for each condition.

dents work with a partner, because students who collaborate with a partner tend to learn more than those who work alone (e.g., Johnson & Johnson, 1994; Webb, 1991). Students also solved practice problems and received mini-lectures during the intervention.

To investigate the role of prior equation solving knowledge, we categorized students as using algebra or not using algebra to solve equations at pretest, and we tested for a Prior Knowledge × Condition interaction. To help bolster our hypothesis that prior knowledge, not general math ability, was the important learner characteristic, we explored whether math ability, as indexed by standardized test scores, interacted with condition.

Materials

Intervention. Packets of worked examples for each condition were adapted from prior work (Rittle-Johnson & Star, 2007, in

press). Three types of equations were used (see Table 1). The three types varied in terms of problem features (defined as characteristics such as number of terms and position of terms with respect to the equals sign). The worked examples illustrated two different solution methods for each type of equation, a conventional method and a shortcut method (see Table 1). For example, the equation $3(x + 5) = 12$ can be solved by first distributing the 3 or first dividing both sides by 3. The latter approach is arguably a shortcut, because it reduces the number of computations and steps needed to solve the equation. All of the shortcut steps relied on treating subexpressions as a composite variable.

The packets were as similar as possible. They all contained eight instances of each of the three equation types for a total of 24 worked examples. Half the worked examples illustrated the conventional solution method, and half illustrated the composite-variable shortcut method. The primary difference between the packets was whether and how the worked examples were paired.

In the compare methods packets, each worked-example pair contained the same equation, solved with the conventional and shortcut methods. In the compare problems packets, each worked-example pair contained two different types of equations, each solved with the same method. For example, a combine composite equation and a divide composite equation were shown together, each solved with the shortcut method (see Figure 1). In the sequential packets, each worked example was presented on a separate page. Across all packets, each solution step was labeled with one of four step labels (distribute, combine, add/subtract on both, multiply/divide on both), because past research indicates that common labels improve the benefits of side-by-side presentation of examples (Namy & Gentner, 2002). Students were asked to complete the labels for most of the steps to encourage active processing of the examples.

Each worked example was accompanied by a question designed to promote reflection on the examples. The questions were designed to exemplify each of Bloom's five levels of thinking (comprehension, application, analysis, synthesis, and evaluation; Bloom, 1956) and were equated across conditions as much as possible. Questions in the compare methods condition focused on comparing the feasibility and efficiency of the solution steps, whereas those in the compare problems condition focused on comparing both the problem features and the particular solution steps (see Figure 1 for examples). Questions in the sequential condition focused on a single example, and some questions focused attention on specific features of the example.

The packets were divided into three sections in order to distribute the material over three intervention sessions. Each section included eight worked examples of two of the three problem types shown in Table 1 and one guided practice problem, on which students were asked to use a particular shortcut method to solve a new equation. At the end of each section, there were four independent practice problems on which students could choose their solution methods. In the compare methods condition, students were asked to solve two practice problems each in two different ways, whereas four different equations were presented in the packets for the other conditions. Three brief homework assignments were developed, each with six problems similar to those solved in class.

Assessment. The same assessment was used as an individual pretest and posttest, and a modified version was used as a retention test. The assessment assessed procedural knowledge, procedural flexibility, and conceptual knowledge and was modified from the one used in Rittle-Johnson and Star (in press). Sample items are included in Table 2.

1. The procedural knowledge measure assessed students' ability to solve equations that had both familiar and unfamiliar problem features. Unfamiliar problem features included additional terms and new operators (i.e., division).

2. Flexibility was measured in two ways, to distinguish knowledge and use of flexible procedures. *Flexible use* was indexed by the frequency of using composite-variable shortcut methods on the procedural knowledge problems. *Flexibility knowledge* was an independent measure of students' knowledge of multiple procedures for solving equations and ability to use these procedures adaptively. Flexibility items fell into three categories: (a) Recognize multiple methods questions (2 items), which assessed students' ability to recognize appropriate first solution steps for a

Table 1
Alternative Solution Methods for Three Types of Equations

Equation type[a]	Sample solution via conventional method	Sample solution via nonconventional shortcut method
$a(x + b) = c$ (divide composite)	$3(x + 1) = 15$ $3x + 3 = 15$ $3x = 12$ $x = 4$	$3(x + 1) = 15$ $x + 1 = 5$ $x = 4$
$a(x + b) + d(x + b) = c$ (combine composite)	$2(x + 1) + 3(x + 1) = 10$ $2x + 2 + 3x + 3 = 10$ $5x + 5 = 10$ $5x = 5$ $x = 1$	$2(x + 1) + 3(x + 1) = 10$ $5(x + 1) = 10$ $x + 1 = 2$ $x = 1$
$a(x + b) = d(x + b) + c$ (subtract composite)	$7(x - 2) = 3(x - 2) + 16$ $7x - 14 = 3x - 6 + 16$ $7x - 14 = 3x + 10$ $4x - 14 = 10$ $4x = 24$ $x = 6$	$7(x - 2) = 3(x - 2) + 16$ $4(x - 2) = 16$ $x - 2 = 4$ $x = 6$

Note. [a] x stands for a variable, and other letters were replaced with numbers.

Table 2
Sample Items for Assessing Procedural, Flexibility, and Conceptual Knowledge

Problem type	Sample items	Scoring
Procedural knowledge		$\alpha = .81$
a. Familiar problem features ($n = 3$)	$\frac{1}{2}(x + 1) = 10$ $3(h + 2) + 4(h + 2) = 35$	1 pt for each correct answer
b. Unfamiliar problem features ($n = 6$)	$3(m - 2)/5 = 33/5$ $3(2x + 3x - 4) + 5(2x + 3x - 4) = 48$	1 pt for each correct answer
Procedural flexibility		
a. Flexible use	Use of composite-variable shortcut method on procedural knowledge problems	$\alpha = .81$
b. Flexibility knowledge		$\alpha = .83$
Generate multiple methods ($n = 4$)	a. Solve this equation in two different ways: $18 = 3(x + 2)$ b. Which of your ways do you think is easiest and fastest?	Part a: 1 pt for two correct, unique solutions Part b: 1 pt for choosing solution with fewest steps
Recognize multiple methods ($n = 2$)	For the equation $2(x + 1) + 4 = 12$, identify all possible steps that could be done next (4 choices).	1 pt for each correct choice
Evaluate nonconventional methods ($n = 2$)	$5(x + 3) + 6 = 5(x + 3) + 2x$ $6 = 2x$ a. What step did the student use to get from the first line to the second line? b. Do you think that this is a good way to start this problem? (a) a very good way; (b) OK to do but not a very good way; (c) not OK to do c. Explain your reasoning.	Part a: 1 pt for correctly identifying step Part b: 2 pt for choice a, 1 pt for choice b, 0 pt for choice c Part c: 2 pt if accurately evaluates efficiency or justifies why OK to do; 1 pt if simply states that step is OK to do
Conceptual knowledge ($n = 12$)		$\alpha = .77$
	1. Which of the following is a like term to (could be combined with) $7(j + 4)$? (a) $7(j + 10)$; (b) $7(p + 4)$; (c) j; (d) $2(j + 4)$; (e) a and d	1. 1 pt for choosing (d)
	2. Here are two equations: $98 = 21x$ $98 + 2(x + 1) = 21x + 2(x + 1)$ (a) Look at this pair of equations. Without solving the equations, decide if these equations are equivalent (have the same answer). (b) Explain your reasoning.	2a. 1 pt for selecting "YES (same answer)" 2b. 1 pt for mentioning equivalence of equations

Note. Cronbach's alphas are for posttest. Alphas were somewhat lower at pretest, largely due to floor effects on some items.

particular problem; (b) Generate multiple methods (4 items), which asked students to solve problems in two different ways or in a different way from a demonstrated solution; and (c) Evaluate nonconventional methods (2 items), which assessed students' ability to evaluate novel solution steps for accuracy and efficiency.

3. There were 12 conceptual knowledge items designed to tap students' verbal and nonverbal knowledge of algebra concepts (e.g., maintaining equivalence and the meaning of variables), sometimes in the context of composite variables. At pretest, students also received four warm-up problems, which were one- and two-step equations. The pretest and posttest were paper-and-pencil assessments.

The retention test was a modified version of the assessment, which could be presented on the computer as part of an ongoing project on computer-assisted problem solving (Razzaq et al., 2005). Students received accuracy feedback after solving each item and had to work on the item until they entered the correct answer; thus, the retention test was more of a dynamic assessment on which students could learn while completing the assessment. Students could not enter explanations or show their work, so explanation and prompted flexibility items were not included on the retention test.

Procedure

Primary data collection occurred within students' intact mathematics classes over five consecutive classroom periods. On Day 1, students were given 45 min to complete the pretest, including 15 min to complete 4 warm-up equations and 9 procedural knowledge items. Some time pressure was included for the procedural items to encourage students to use efficient solution methods.

On Day 2, each student was randomly paired with another student in the class, leading to 114 groups; in 8 of these groups, there were 3 students due to uneven numbers of students in a class. The day began with a 10-min, scripted whole-class lesson. First, students attempted to solve the equation $3(x + 1) = 12$ on their own. The instructor then worked through the conventional solution to the problem, labeling each step. The instructor concluded with a model of appropriate partner work to show students how to work through the packets in pairs (modeling discussion of identifying and labeling the steps in a worked example and answering a simple reflection prompt).

Following this introduction, student pairs began working on the packets. For the worked examples, they were instructed to describe each solution method to their partner and discuss the accompany-

ing questions before writing down their answer. For the practice problems, students were asked to solve the problem on their own, compare answers with their partner, and have their answers checked by an adult. The classroom teacher and one or two members of the project team circulated through the class; they provided help with implementing steps but not with choosing solution steps or answering reflection questions. Students were given the same homework assignment at the end of the class period.

The next 2 days of instruction followed the same format, with a brief whole-class lesson introducing a new problem feature (variables on both sides on Day 3 and fractional coefficients on Day 4) through demonstration of how to solve a sample problem using the conventional method, followed by partner work on the packets for the day. Students started a new packet each day and did not return to finish incomplete packets from the previous day. At the end of Day 4, the instructor provided an 8-min wrap-up lesson that emphasized (a) there is more than one way to solve an equation, (b) any way is OK if the two sides of the equation are kept equal, and (c) some ways of solving equations are better or easier than others. These points were mentioned throughout previous lessons, and we included a summary lesson because direct instruction can augment the benefits of comparison (Schwartz & Bransford, 1998; VanderStoep & Seifert, 1993).

On Day 5, students were given 45 min to complete the posttest, which was identical in content and administration to the pretest, except that the four warm-up equations were not included. At least 2 weeks later, students completed the retention test as part of their regularly scheduled time in the computer lab. The retention test was presented on the computer and was administered by the classroom teacher. Students were to be given 45 min to complete the retention test, although some technical difficulties with the computers during some class periods led to less time spent on the assessment.

Coding

Assessment. The nine equations on the procedural knowledge assessment were scored for accuracy of the answer, and students received a percentage correct based on the number of problems solved correctly out of nine. Solution methods were also coded, based on whether the first step (a) was distributed across parentheses (the conventional solution method), (b) was a shortcut step that had been demonstrated in the worked examples (see Table 1), (c) was an unusual or incorrect algebraic step, or (d) used an informal, nonalgebraic approach (e.g., guess-and-test or unwind) or whether the student (e) did not attempt the problem. We used whether students used algebraic methods (the first three codes) at pretest as our index of prior knowledge of algebraic methods. Frequency of using a shortcut method at posttest was used as an indicator of flexible use of procedures, because it is both a nonconventional and a more efficient way to solve the target equations.

The flexibility and conceptual knowledge assessment was scored according to guidelines in Table 2, and totals were converted to percentage correct. Independent coders coded the solution methods and explanation qualities across the assessment for 20% of the sample, and exact agreement ranged from 90% to 99%. Discrepancies were discussed, and codes were altered when deemed appropriate by the primary coder.

Intervention. For each student, we tallied how many practice problems were completed and how often the demonstrated shortcut was used. Interrater reliability on 20% of the sample for use of the shortcut method was 98%.

We also coded students' written explanations on the basis of the different types of comparisons made and general features of the explanations. The codes are described in the Results section. Exact agreement on presence of each explanation type, conducted by two raters on 20% of the sample, ranged from 91% to 96%.

Data Analysis

Missing data. Three students were absent at pretest, and 9 different students were absent at posttest. The amount of missing data was much higher for the retention test, which was administered by teachers; 46 students did not take the retention test. Another 43 students began the assessment but did not have time to begin the conceptual knowledge portion because it was presented last. Often this was due to technical difficulties in the computer lab, which reduced the amount of time students had to work on the assessment during the class period.

We imputed missing data because simulation studies indicate that imputation leads to the same conclusions as when there are no missing data, if the data are missing at random and less than 20% of the data are missing (Barzi & Woodward, 2004; Schafer & Graham, 2002). Little's MCAR test confirmed that our data were missing completely at random, $\chi^2(476, N = 236) = 35.219$, $p = 1.0$, and data were missing on less than 20% of cases for all measures but one. On the conceptual knowledge measure at retention test, 37% of cases were missing, but we decided to impute the missing data because there were not better alternatives. We interpret those findings with caution.

To impute the missing data, we used the expectation-maximization algorithm for maximum-likelihood estimation via the missing value analysis module of SPSS, as recommended by Schafer and Graham (2002). The students' missing scores were estimated from all nonmissing values on continuous variables that were included in the analyses presented below. Findings obtained using a casewise deletion approach yielded the same basic findings.

Multilevel models. Children worked with a partner for the intervention, so we calculated intraclass correlations to test for nonindependence in partner scores on the posttest, controlling for the predictor variables (Kenny, Kashy, & Cook, 2006). Intraclass correlations ranged from −.05 to .31, and there were significant intraclass correlations on multiple measures. Because traditional analysis of variance models assume independence in the data, we used multilevel linear modeling to account for this nesting within dyad. We specified the use of restricted maximum-likelihood estimation and compound symmetry for the variance–covariance structure in the models (Kenny et al., 2006). The significance tests used the Satterthwaite (1946) approximation to estimate the degrees of freedom.

Because students worked in pairs, our model had two levels: the individual level and the dyad level. Effects of pretest knowledge were tested in the individual level of the model (e.g., whether students used algebra on the pretest and their pretest conceptual, procedural, and

flexibility knowledge scores). Effects of partners' pretest knowledge were also tested at the individual level, in line with Kenny's actor–partner interdependence model (see http://davidakenny.net/dyad.htm for a tutorial and details on implementing this approach in SPSS). In other words, both a person's own scores and the partner's scores were explored as predictors of the individual's outcomes in the first level of the model. Effects of experimental condition were tested in the second level, as were class type (regular or honors) and grade level (seventh or eighth). To test for an aptitude–treatment interaction, we included a cross-level interaction term between condition and use of algebra at pretest. To help support our choice of algebra use at pretest as our "aptitude" measure, we also explored whether an interaction term between condition and students' standardized test score on the Measures of Academic Progress was significant.

Initial analyses were conducted to confirm the appropriate choice of control variables to include in the model. Class type (regular or honors) and pretest knowledge on the different measures almost always influenced performance, so each was included in all of the models. Partners' knowledge at pretest predicted performance on the procedural knowledge assessment at posttest but not the other measures. Grade level did not impact posttest performance, but it did impact retention test performance. Also, participants' gender did not impact any outcome. For parsimony, only control variables that impacted performance on a particular measure were included.

Results

First, we overview students' knowledge at pretest. Next, we report the effects of condition and students' use of algebra at pretest on knowledge at posttest and at retention test. Finally, we explore how condition affected intervention activities, such as the characteristics of students' explanations.

Pretest Knowledge

At pretest, students had some procedural, flexibility, and conceptual knowledge (see Table 3). However, students had much lower knowledge of equation-solving methods than did students in our past studies. Only 20% used the distribute-first step for solving equations at least once, and it was used on only 8% of trials. Another 39% of students attempted to use algebra at least once but often used it incorrectly, such as by subtracting a value twice on the same side of an equation (used on 15% of trials). A few of these students (4%) also used shortcut methods. The remaining 41% of students never used an algebraic approach to solve an equation at pretest. As our index of prior knowledge, we categorized students as using algebra at least once, whether correctly or incorrectly, or as never using algebra (59% vs. 41% of students). Students who attempted algebraic methods were familiar with equation-solving methods, but they certainly were not proficient equation solvers. Compared with students who did not use algebra, they were only marginally more likely to be in honors sections (75% vs. 64% of students), $\chi^2(1, N = 236) = 2.943$, $p = .086$, or to be in the eighth grade (85% vs. 76%), $\chi^2(1, N = 236) = 3.943$, $p = .074$.

There were no significant differences between conditions in procedural, flexibility, and conceptual knowledge, but students in the compare problems condition were less likely to use algebra to solve an equation at least once at pretest than were students in the compare methods or sequential conditions (45% vs. 61% and 69% of students, respectively), $\chi^2(2, N = 236) = 9.914$, $p = .007$.

Overall, students in this study had less knowledge of equation-solving procedures than did students in previous studies, and there were not significant differences in pretest knowledge between conditions.

Effects of Condition and Prior Knowledge on Posttest Performance

Condition did not have a large impact on performance by itself (see Table 3) but rather depended on students' use of algebra at pretest (see Figure 2). Table 4 shows the results of the two-level

Table 3
Performance on Outcome Measures by Condition

	Pretest		Posttest		Retention test	
Outcome	*M*	*SD*	*M*	*SD*	*M*	*SD*
Procedural knowledge (% correct)						
Compare methods	17.9	20.5	39.1	32.4	58.4	32.7
Compare problems	19.8	17.1	40.2	29.9	67.8	24.1
Sequential	21.9	21.5	44.8	26.5	64.0	25.8
Flexible use (% trials used shortcut method)						
Compare methods	0.1	1.2	22.7	30.7	na	na
Compare problems	0.7	3.3	22.4	28.6	na	na
Sequential	1.9	8.1	25.6	27.5	na	na
Flexibility knowledge (% correct)						
Compare methods	27.1	16.7	55.3	25.3	49.0	20.8
Compare problems	29.7	17.1	54.1	20.9	53.6	19.1
Sequential	29.8	18.5	61.3	21.5	51.4	18.7
Conceptual knowledge (% correct)						
Compare methods	24.1	20.7	42.5	25.9	33.3	26.5
Compare problems	24.0	18.4	52.4	22.4	38.9	23.2
Sequential	23.8	21.2	51.1	22.4	35.4	24.0

Note. Students could not show their work on the retention test, so only a subset of items from the pretest and posttest was included. na indicates that flexible use could not be coded at retention test because students could not show their solution methods on the computerized assessment.

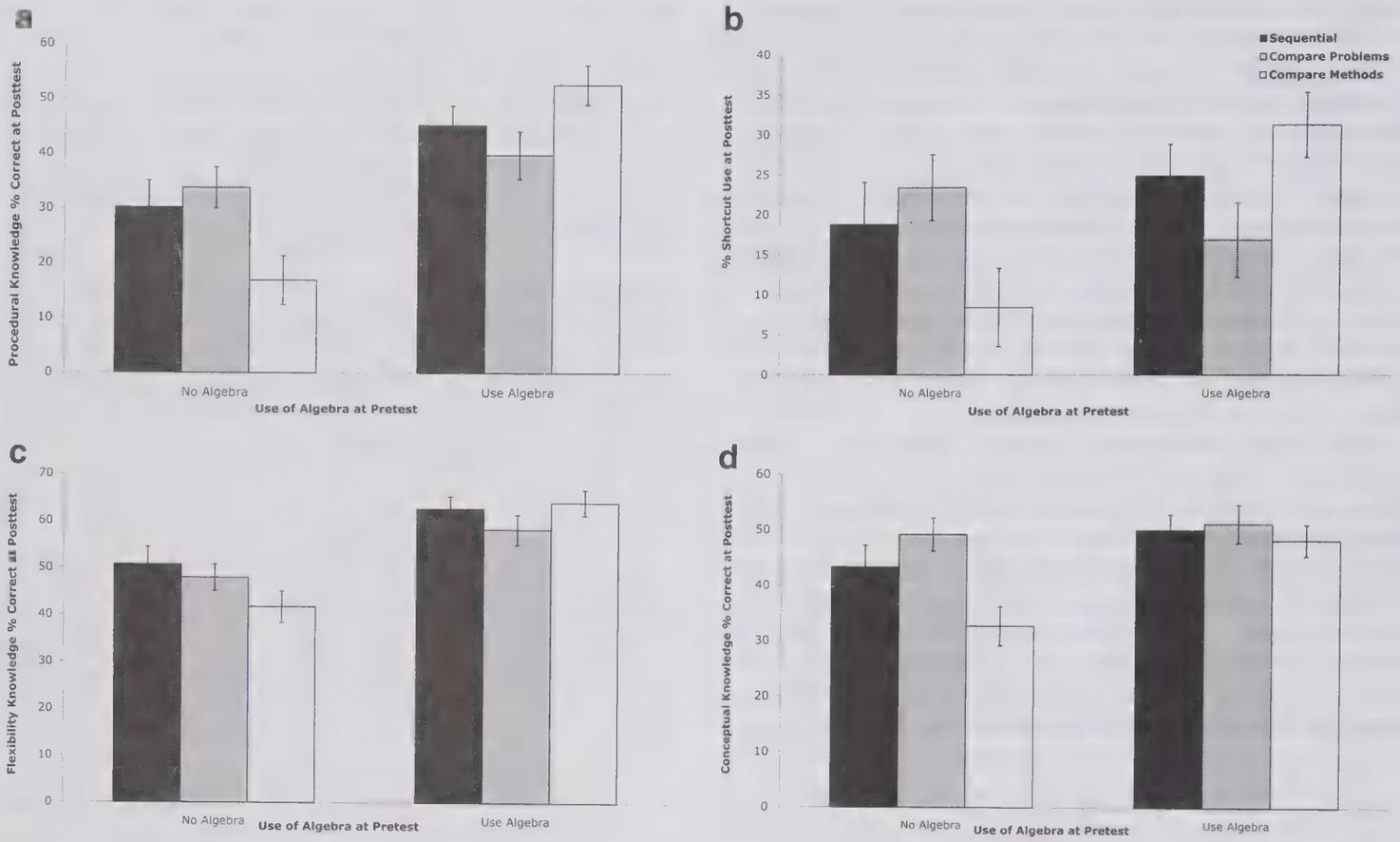


Figure 2. Effects of condition and pretest use of algebra on (a) procedural knowledge, (b) flexible use of procedures, (c) flexibility knowledge, and (d) conceptual knowledge. Values are estimated marginal means with standard error bars.

linear modeling analyses predicting procedural knowledge (col. 1), flexible use of procedures (col. 2), flexibility knowledge (col. 3), and conceptual knowledge (col. 4). As described in the Method section, the models included condition, use of algebra at pretest, the interaction between the two, and several control variables.

We explored whether students' mathematics ability rather than students' algebra use at pretest was a more appropriate predictor of the effectiveness of instructional condition. We conducted analyses parallel to those reported below but used standardized math test score, rather than use of algebra, as a main effect and an interaction term (with condition). Standardized math test score did not interact with condition in predicting any of the outcomes.

Procedural knowledge. The effect of condition depended on whether students used algebra at pretest. There was not an overall main effect for condition, but there was a main effect for algebra use at pretest and a Condition × Algebra Use interaction, $F(2, 108) = 0.252$, $p = .778$; $F(1, 215) = 34.736$, $p < .001$; and $F(2, 213) = 8.466$, $p < .001$, respectively. As shown in panel a of Figure 2, comparing methods led to lower procedural knowledge if students did not use algebra at pretest but to higher procedural knowledge if students did use algebra at pretest. The effects of comparing problem types were more modest and did not differ significantly from the sequential condition or by prior algebra use.

Parameter estimates from the models, shown in Table 4, confirmed these effects. Not using algebra at pretest and the sequential condition were the reference categories in the models. For students who did not use algebra at pretest, comparing methods led to a 13% decrease in accuracy relative to the sequential condition ($\beta = -13.3$, $p = .043$; see Table 4, column 1), whereas comparing problem types led to a nonsignificant increase in accuracy ($\beta = 3.5$, $p = .563$). For students who did use algebra at pretest, this effect of condition was offset by an interaction term. Relative to those who did not use algebra, students who used algebra at pretest scored 21 points higher in the compare methods condition ($\beta = 20.6$, $p = .008$) but a nonsignificant 9 points lower in the compare problems condition ($\beta = -9.0$, $p = .232$).

Thus, comparing methods increased procedural knowledge for students who used algebra at pretest but decreased procedural knowledge for students who did not use algebra at pretest. The effects on procedural knowledge of comparing problems did not differ significantly from the sequential condition.

Flexible use of solution methods. Condition and prior knowledge also influenced flexibility in procedure use on the procedural knowledge items. On these problems, the composite-variable shortcut was more efficient than a conventional, distribute-first method and thus indicated more adaptive and flexible use of solution methods. Students rarely used other valid methods to solve the equations, with less than 1% of trials being solved correctly using other algebraic methods and only 3% of trials being solved correctly via nonalgebraic methods.

Frequency of using shortcuts depended both on condition and on whether students used algebra at pretest. There was not a main

Table 4
Parameter Estimates for Posttest Outcomes

Variable	Procedural knowledge	Flexible use	Flexibility knowledge	Conceptual knowledge
Intercept	25.67 (5.84)***	14.88 (5.90)*	47.84 (4.05)***	38.41 (4.29)***
Condition (reference = sequential)				
Compare methods	−13.29 (6.53)*	−10.39 (7.08)	−9.02 (4.92)	−10.64 (5.23)*
Compare problems	3.53 (6.10)	4.59 (6.58)	−2.82 (4.58)	5.85 (4.87)
Pretest algebra use (reference = no algebra)	14.93 (5.60)**	5.94 (5.87)	11.91 (4.33)**	6.61 (4.70)
Condition × Algebra Use				
Compare methods	20.63 (7.64)**	16.72 (8.08)*	10.28 (5.90)	8.77 (6.39)
Compare problems	−8.95 (7.46)	−12.54 (7.79)	−1.62 (5.78)	−4.74 (6.31)
In honors class	13.53 (3.90)***	7.81 (4.29)	5.78 (2.80)*	9.77 (2.89)***
Pretest conceptual	0.37 (0.80)***	0.41 (0.09)***	0.27 (0.06)***	0.38 (0.07)***
Pretest procedural	0.47 (0.08)***	0.25 (0.08)**	0.33 (0.06)***	0.32 (0.07)***
Pretest flexibility	0.26 (0.10)***	0.32 (0.10)***	0.29 (0.07)***	0.26 (0.08)***
Partner conceptual	−0.10 (0.08)	na	na	na
Partner procedural	−0.05 (0.09)	na	na	na
Partner flexibility	0.30 (0.10)**	na	na	na
Partner used algebra	−3.77 (3.23)	na	na	na

Note. $N = 236$. Unstandardized coefficients are shown with standard errors in parentheses. Pretest conceptual, procedural, and flexibility knowledge were grand mean centered. na indicates that the variable was not included in the final model because preliminary analyses indicated that it was not a reliable predictor.
* $p < .05$. ** $p < .01$. *** $p < .001$.

effect for condition, but there was a main effect for algebra use and a Condition × Algebra Use interaction, $F(2, 115) = 0.098$, $p = .906$; $F(1, 214) = 4.728$, $p = .031$; and $F(2, 210) = 7.292$, $p < .001$, respectively. As shown in Figure 2, panel c, comparing methods tended to decrease flexible use of procedures if students did not use algebra at pretest but to increase flexible use for those who did. The effects of comparing problem types were very modest but tended to be opposite of the effects of comparing methods.

Parameter estimates from the model confirmed these findings (see Table 4, column 2). For students who did not use algebra at pretest, comparing methods led to a 10-point decrease in accuracy compared to the sequential condition ($\beta = -10.4$, $p = .144$), whereas comparing problem types led to a very modest 5-point increase ($\beta = 4.6$, $p = .486$); these effects did not reach significance. For students who did use algebra at pretest, the effect of condition was offset by an interaction term. Relative to those who did not use algebra, students who used algebra at pretest scored 17 points higher in the compare solution methods condition ($\beta = 16.7$, $p = .040$) and a nonsignificant 13 points lower in the compare problems condition ($\beta = -13.5$, $p = .109$).

Finally, there was a benefit to using shortcut methods: their use improved accuracy. Frequency of using shortcuts at posttest was positively related to accuracy at posttest, $r(230) = .58$, $p < .001$, after controlling for pretest knowledge measures.

Flexibility knowledge. Next, we consider knowledge of multiple procedures and when to use them. The findings were consistent with the other measures, but the effects were more marginal. There was not a main effect for condition, but there was a main effect for algebra use at pretest and a marginal Condition × Algebra Use interaction, $F(2, 117) = 0.921$, $p = .401$; $F(1, 226) = 35.888$, $p < .001$; and $F(2, 224) = 2.548$, $p = .080$, respectively. As shown in Figure 2, panel c, comparing methods tended to be less beneficial for flexibility knowledge than other conditions if students did not use algebra at pretest; otherwise, condition had minimal influence on flexibility knowledge.

Parameter estimates confirmed this interpretation (see Table 4, column 3). For students who did not use algebra at pretest, comparing methods led to a marginal 9-point decrease in accuracy ($\beta = -9.0$, $p = .068$), and comparing problem types led to a nonsignificant 3-point decrease in accuracy ($\beta = -2.82$, $p = .538$). For students who did use algebra at pretest, this effect of condition was offset by an interaction term. Relative to those who did not use algebra, students who used algebra at pretest scored 10 points higher in the compare solution methods condition ($\beta = 10.3$, $p = .083$) but about the same in the compare problems condition ($\beta = -1.6$, $p = .779$).

We conducted follow-up analyses on the three subscales of the flexibility knowledge measure. The effects of condition and algebra use at pretest were the same for the Generating Multiple Methods and Evaluating Nonconventional Methods subscales. Condition had minimal impact on recognizing multiple methods. Overall, condition had limited impact on flexibility knowledge, but trends were in line with findings on other measures.

Conceptual knowledge. Finally, consider students' knowledge of related concepts, such as equivalence, like terms, and composite variables. There was a main effect for condition, a main effect for algebra use at pretest, and a marginal interaction between the two, $F(2, 116) = 4.803$, $p = .010$; $F(1, 226) = 8.827$, $p = .003$; and $F(2, 226) = 2.497$, $p = .085$, respectively. As shown in Figure 2, panel d, comparing methods tended to be less supportive of conceptual knowledge than did other conditions if students did not use algebra at pretest; otherwise, condition had minimal influence on conceptual knowledge.

Again, parameter estimates from the model supported this interpretation of the omnibus tests (see Table 4, column 4). For students who did not use algebra at pretest, comparing methods led to an 11-point decrease in conceptual knowledge scores ($\beta = -10.6$, $p = .043$) and comparing problem types led to a nonsignificant 6-point increase ($\beta = 5.9$, $p = .231$). Students who used algebra at pretest, relative to those who did not use algebra, scored about 9 points higher in the compare methods condition ($\beta = 8.8$,

$p = .171$) and about 5 points lower in the compare problems condition ($\beta = -4.7$, $p = .453$). Overall, students who did not use algebra at pretest had lower conceptual knowledge in the compare methods condition than in the other two conditions. For students who did use algebra, effect of condition was limited.

Teacher effects. We also explored whether the effect of condition depended on students' teachers. The teachers may have varied in how much they encouraged use and comparison of multiple solution methods or how much they used comparison in general, and this prior experience could impact students' readiness to learn from our different instructional conditions. Inspection of means in each condition for the five teachers suggested that students of one of the teachers consistently learned more in the compare methods condition than in the other conditions, whereas there was less of an effect of condition for the other teachers. Statistical models that included teacher and a Teacher × Condition interaction term, in place of the algebra use at pretest, supported the inference that one of the teachers had better prepared his students to learn by comparing methods. There was a significant effect of being in this teacher's classes for performance in the compare methods condition on three of the four outcome measures (procedural knowledge, flexible use, flexibility knowledge). Prior to our intervention, we did observe this teacher challenging students to find the more efficient method for solving a "problem of the day" (that was not algebraic), something we did not observe in the other four classrooms. However, these findings must be interpreted with great caution, given the relatively small sample size for exploring teacher effects and our very limited observations of teaching.

Effects of other prior knowledge measures. In all four models, pretest knowledge on each measure was a strong predictor of posttest performance (see Table 4). This result supports the mutually important and beneficial roles of each type of knowledge—conceptual, procedural, and flexibility—for knowledge acquisition.

How much a student's partner knew at pretest was much less influential. As noted in the *Data Analysis* section, partners' pretest scores did not influence performance on the flexible use, flexibility knowledge, or conceptual knowledge measures, so they were not included in the final models for those outcomes. However, partners' pretest knowledge did predict accuracy on the procedural knowledge measure. In particular, partners' flexibility knowledge positively predicted procedural knowledge ($\beta = 0.30$, $p = .003$), although partners' conceptual and procedural knowledge or use of algebra at pretest did not. It may be that openness and awareness of alternative solution methods is an important characteristic for students' conversational partners to have when studying multiple solution methods. Somewhat surprisingly, partners' pretest flexibility did not impact flexible procedure use or flexibility knowledge.

Effects of Condition and Prior Knowledge on the Retention Test

The retention test tapped transfer of knowledge to a different assessment format (paper-and-pencil vs. computer) and environment (administered by teacher in computer lab rather than by researchers in the classroom) after a 2-week delay. The computer system provided accuracy feedback after each item, which gave students some opportunity to learn while taking the assessment. Many students ran out of time before completing the assessment, so accuracy scores are based on the number of items each child completed. Because so many students missed or did not complete the retention test, these findings must be interpreted with caution. However, for the most part, they are consistent with the posttest findings. This suggests that the effects of condition and prior knowledge were stable. Analyses of retention test performance paralleled analyses of posttest performance, except that (a) grade level impacted retention performance on some measures in preliminary analyses and so was included as a covariate in the models and (b) partners' scores did not impact performance on any measure and so were not included in the models.

Procedural knowledge. For procedural knowledge, there was an overall main effect for condition and for algebra use at pretest and a marginal Condition × Algebra Use interaction, $F(2, 103) = 6.075$, $p = .003$; $F(1, 217) = 20.923$, $p < .001$; and $F(2, 211) = 2.451$, $p = .089$. For students who did not use algebra at pretest, comparing methods led to a significant 14-point decrease in accuracy relative to the sequential condition ($\beta = -13.8$, $p = .041$), whereas comparing problem types led to a nonsignificant 7-point increase ($\beta = 6.9$, $p = .283$). Students who used algebra at pretest, relative to those who did not use algebra, scored about 17 points higher in the compare methods condition ($\beta = 16.7$, $p = .048$) and about the same in the compare problems condition ($\beta = 1.9$, $p = .818$). Pretest conceptual and procedural knowledge also positively predicted performance. Thus, on the retention test and similar to posttest, students who did not use algebra had lower procedural knowledge in the compare methods condition than in the other conditions. Students who used algebra at pretest performed best when they had compared methods.

Flexibility knowledge. Students could not show their work while working on the computer system, so we could not gather data on flexible use of procedures or on ability to generate multiple procedures when prompted (i.e., prompted flexibility). Explanation items on the Evaluating Flexibility subscale also could not be presented, so most of the items on the flexibility retention test were recognize flexibility items. Given that the conditions did not differ on this subscale at posttest, it was not surprising that there was no main effect for condition or Condition × Algebra Use interaction, $F(2, 116) = 1.613$, $p = .204$, and $F(2, 224) = 0.962$, $p = .384$, respectively.

Conceptual knowledge. Recall that 37% of students did not begin this assessment. We imputed the missing data, but with this high level of missing data, all effects must be interpreted cautiously. There was a main effect of condition but no interaction with algebra use at pretest, $F(2, 108) = 5.837$, $p = .004$, and $F(2, 223) = 0.943$, $p = .391$, respectively. For students who did not use algebra at pretest, there was no effect of comparing methods relative to sequential study of examples ($\beta = 0.9$, $p = .865$), but comparing problem types led to an almost 13-point increase in scores ($\beta = 12.7$, $p = .008$). For students who did use algebra at pretest, the effects of condition were not offset much ($\beta = 4.0$ for compare methods and $\beta = -4.1$ for compare problems; $ps > .5$). Comparing problem types may have aided retention of conceptual knowledge.

Effects of Condition on Intervention Activities

To help understand how condition and use of algebra at pretest impacted learning, we examined students' responses during the

intervention. Recall that students studied packets of worked examples, answered explanation prompts about the examples, and solved practice problems.

Practice problems. Students solved an average of 10 of the 12 independent practice problems (M = 9.2, 9.5, and 10.4 of attempted problems for the compare methods, compare problems, and sequential conditions, respectively). We conducted a two-level linear model, with use of algebra at pretest at Level 1, condition at Level 2, and the interaction between the two as a cross-level interaction. To conserve space, we simply report parameter estimates from the model. Students who did not use algebra at pretest solved 2 fewer problems in the compare methods condition than in the sequential condition ($\beta = -1.7$, $p = .045$). Students who did use algebra solved an additional problem in the compare methods condition, but this effect was not significant ($\beta = 0.8$, $p = .341$). Students in the compare problems condition solved the same number of problems as did those in the sequential condition, regardless of pretest algebra use. Thus, the compare methods condition may have been more difficult for students who did not use algebra at pretest, leading them to solve fewer problems. However, this was not true for students who did use algebra at pretest.

Students chose to use the more efficient, shortcut method on 32% of attempted practice problems, and frequency of use did not vary significantly by condition (M = 25%, 38%, and 33% of attempted problems for the compare methods, compare problems, and sequential conditions, respectively) or by pretest use of algebra. Students were also prompted to implement a shortcut method on three guided practice problems. Success on the guided practice problems did vary by condition (M = 2.1, 2.2, and 2.5 out of 3 problems, respectively), particularly for students who did not use algebra at pretest. For those who did not use algebra, students who compared solution methods were less able to implement the shortcut method correctly ($\beta = -0.6$, $p = .011$) than were students who studied the examples sequentially. Students who used algebra had a bit more success in the compare methods condition than did those who had not used algebra ($\beta = 0.4$, $p = .130$). Students who compared problem types solved the same number of problems as did those in the sequential condition, regardless of pretest algebra use. These findings suggest that students in the compare methods condition, especially those who did not use algebra at pretest, were struggling to solve the practice problems and were not becoming proficient in use of shortcut methods.

Explanations on worked examples. A majority of the intervention was spent studying and explaining worked examples. Students answered 74% of the 24 available explanation prompts, but this varied by condition (M = 70%, 75%, and 77% of questions answered for the compare methods, compare problems, and sequential conditions, respectively), particularly for students who did not use algebra at pretest. For students who did not use algebra at pretest, those in the compare methods condition answered 13% fewer questions than did those in the sequential condition ($\beta = -12.7$, $p = .010$). Students who used algebra at pretest, relative to those who did not use algebra, answered 10% more of the questions in the compare methods condition ($\beta = 10.2$, $p = .018$). There was no effect for comparing problems, regardless of algebra use at pretest. Comparing methods appeared to slow down students who did not use algebra.

The explanation prompts were designed to facilitate the appropriate processes for each condition, and thus they varied by condition. Because students discussed their explanations with their partner, we coded the written explanations of a randomly selected member of each pair. This eliminated concerns of nonindependence in the data, so we used analysis of covariance models. The explanation coding scheme was adapted from previous work and included explanation features that were found to be important in past studies. The dependent measure in all analyses was based on the number of questions answered for each child to help equate for any differences in the number of questions answered. Due to the exploratory nature of these analyses, which required the use of multiple tests, we adopted the more conservative alpha value of .005 when interpreting the findings.

First, consider the four types of comparisons students made: the efficiency of the methods, specific solution steps, problem features, and answers (see Table 5). As intended, students in the compare methods and compare problems conditions made comparisons almost three times as often as did those in the sequential condition. Students in both comparison conditions were most likely to compare solution steps and did so on about half of their explanations. Students in the compare methods condition also compared the efficiency of the methods, and students in the compare problems condition were more likely to compare problem features. Students in the sequential condition also compared problem features (e.g., when asked whether a demonstrated method could be used to solve a different equation). Students rarely compared answers (2% of trials). As intended, both comparison conditions focused attention on individual solution steps, but the two focused attention differentially on comparing efficiency or problem features. Some reflection prompts in the sequential condition also encouraged comparison of problem features.

We also coded general characteristics of students' explanations, including references to multiple methods, mention of the shortcut step, evaluations of the examples, and use of mathematical terminology, as shown in Table 5. In a result parallel to findings on specific comparisons, students in the compare methods condition were most likely to reference multiple solution methods and to evaluate the efficiency of methods. Students in the sequential condition were most likely to use mathematical terminology and were least likely to note shortcut steps or reference multiple methods. Across conditions, students rarely evaluated problem features or accuracy. It is worth noting that use of algebra at pretest had very little effect on explanation qualities, with the exception that students who used algebra at pretest were more likely to evaluate problem features, $F(1, 107) = 6.632$, $p = .011$, $\eta^2 = .058$.

In addition, we evaluated which explanation features predicted performance on the posttest to explore which types of explanations led to better learning outcomes. In the models, frequency of each of the explanation types was used as a predictor of procedural knowledge, flexible use, flexibility knowledge, or conceptual knowledge at posttest. Pretest knowledge measures, algebra use at pretest, and class type were included as covariates. Note that only half the sample was included in these analyses, because we coded explanation quality for only 1 member of each pair.

Unlike in past studies, comparing solution steps or efficiency did not predict any of the outcomes. Rather, comparing problem features positively predicted procedural knowledge, $F(1, 105) =$

Table 5
Percentage of Intervention Explanations Containing Each Feature, by Condition

Explanation characteristic	Sample explanations	Compare methods	Compare problems	Sequential
1. Any comparison[a]	At least one comparison	74	77	28
Compare efficiency of methods[b]	"[I would use] Erica's way because it has less steps."	19	9	1
Compare solution steps[b]	"Jessica distributed and Mary combined like terms," "They both subtracted 4."	49	55	8
Compare problem features[c]	"Patrick has a variable on both sides of the equation and Abby has them both on one side."	8	19	20
2. Reference multiple methods[b]	"It is okay to do it either way."	90	78	17
3. Note shortcut step[a]	"Patrick subtracted $3(y + 1)$ first."	42	37	26
4. Evaluate:				
Efficiency[c]	"Mary's way is more compact and might take less time."	42	21	21
Problem features	"Heather's problem has easier numbers."	7	7	7
5. Mathematical terminology[a]	"They got variables on *both sides*."	12	19	30

Note. [a] Sequential differs from other two conditions at $p \leq .005$. [b] All three conditions differ from each other at $p < .005$. [c] Compare methods differs from other two conditions at $p < .001$.

3.791, $p = .054$, $\eta^2 = .035$; flexible procedure use, $F(1, 105) = 7.488$, $p = .007$, $\eta^2 = .070$; and conceptual knowledge, $F(1, 105) = 4.05$, $p = .047$, $\eta^2 = .037$. Frequency of different types of comparisons did not predict flexibility knowledge, but two general explanation characteristics did: evaluating problem features, $F(1, 105) = 4.585$, $p = .035$, $\eta^2 = .042$, and justifying answers with mathematical terminology, $F(1, 105) = 3.868$, $p = .052$, $\eta^2 = .036$. Evaluating problem features also predicted flexible use, $F(1, 105) = 13.356$, $p < .001$, $\eta^2 = .118$, but not the other two outcomes. Overall, focusing on problem features seemed to help students learn in this study, regardless of condition.

Summary

Students began the study with varied but fairly limited knowledge of algebraic methods for solving multistep linear equations. Of the students, only 20% used a correct algebraic method to solve at least one pretest equation, 40% attempted to use algebra but did so incorrectly, and the remaining 40% did not attempt to use algebraic methods. In turn, whether students attempted to use algebra at pretest influenced which instructional condition was most effective at supporting student learning.

Students who did not use algebra at pretest had higher performance at posttest if they studied examples sequentially or compared two problem types solved in the same way. Students were able to complete more of the assessment materials in these conditions and spent more time focused on problem features. In turn, frequency of comparing or evaluating problem features predicted posttest performance across measures. Students in these two conditions also had less difficulty implementing the shortcut method during the intervention.

The negative effects of comparing methods were not present for students who attempted to use algebra at pretest. Across measures, comparing methods was more effective for students who used algebra at pretest than for students who did not. In contrast, the effects of comparing problem types were not dependent on algebra use at pretest.

Discussion

As expected, learners' prior knowledge of equation solving altered the effectiveness of comparison relative to sequential study of the same material. In particular, students who did not attempt algebraic methods at pretest learned less if they compared solution methods. Regardless of whether students attempted to use algebraic methods, comparing problem types had limited impact on outcomes. In the discussion, we first integrate the current findings with previous findings on the effects of comparison for mathematics learning and on aptitude–treatment interactions. Next, we discuss what, when, and how comparison is effective and the implications for theories of analogical learning. Finally, we consider the educational implications for teaching children who are novices in a domain.

Integrating Past Results on Comparison

Descriptive studies of mathematics teaching have indicated that many teachers, especially expert teachers, use comparison in their lessons (e.g., Ball, 1993; Lampert, 1990; Richland et al., 2007). Reform efforts for mathematics education have championed a particular type of comparison: comparing solution methods (NCTM, 2000). In support of this position, three recent studies have found that students who were randomly assigned to compare solution methods learned more than those who were not so assigned across a variety of outcome measures and in two very different domains (see Table 6). Comparing solution methods has been particularly effective at supporting flexibility knowledge and flexible procedure use across studies, as well as supporting conceptual and procedural knowledge in some studies. In these studies, most students knew one of the solution methods at pretest. In contrast, students in the current study often did not, and this low prior knowledge of solution methods mattered.

Students' work during the intervention, and how it related to posttest performance, offered some clues as to why comparing methods was less supportive of learning for students who did not use algebraic methods at pretest. These students completed less of the intervention materials and struggled to use nontraditional, shortcut methods correctly. They did make frequent comparisons, often of the particular solution steps or the efficiency of the methods. However, unlike in past studies, these types of comparative explanations were not predictive of learning. In past studies, acceptance and use of multiple procedures was a major benefit of comparing solution methods, but until students can comprehend

Table 6
Summary of Features and Outcomes From Experimental Studies on Using Comparison to Support Mathematics Learning

Feature	Rittle-Johnson & Star (2007)	Rittle-Johnson & Star (in press)	Star & Rittle-Johnson (2009)	Current study
Instructional conditions	Compare methods or sequential	Compare methods, compare problems, or compare equivalent[a]	Compare methods or sequential	Compare methods, compare problems, or sequential
Target task	Linear equations	Linear equations	Computational estimation	Linear equations
% children familiar with a target method at pretest	96%	69%	79%	20% (59% with broader criteria)
Condition with highest performance				
Conceptual knowledge	Same (poor measure)	Compare methods	Depends on prior knowledge	Depends on prior knowledge
Procedural knowledge	Compare methods	Same	Same	Depends on prior knowledge
Flexibility knowledge	Compare methods	Compare methods	Compare methods	Depends on prior knowledge
Flexible use	Compare methods	Compare methods and compare problems	Compare methods	Depends on prior knowledge

Note. [a] In the compare equivalent condition, students compared equivalent equations solved with the same solution method. The equations varied only in minor surface features, for example, $2(x + 3) = 8$ and $5(y + 4) = 10$.

multiple procedures, it may be of limited value for them to compare the benefits and drawbacks of each procedure with a partner.

Synthesizing across our studies on comparing solution methods, we note that the pattern of results suggests that comparing methods, rather than sequential study, can be harmful for students with low prior knowledge of solution methods; neutral for students who are attempting to master one of the to-be-compared methods; and beneficial for students who accurately use one of the to-be-compared methods. Note that these findings are in the context of minimal teacher support; it may be that comparing solution methods is effective for novices if more instructional support is provided.

Comparing solution methods is the type of comparison typically used in mathematics education, but comparing problem types is another useful type of comparison to explore. It is the dominant form of comparison used in laboratory experiments in cognitive science, although these experiments have not used mathematical tasks, included a range of types of learning outcomes, or been conducted with school-age children. In the current study, the comparing problems condition was not more (or less) effective than the sequential condition. This may be because the two conditions were equally likely to elicit comparison and evaluation of problem features during the intervention, and the frequency of these two types of explanations was a positive predictor of posttest performance. Indeed, past research suggests that focusing attention on critical problem features supports knowledge transfer (Bransford, Franks, Vye, & Sherwood, 1989) and procedural flexibility (Rittle-Johnson & Star, in press; see Table 6). In addition, prior knowledge had little effect on learning in the compare problems condition. Together, these findings suggest that comparing problem features is useful for supporting mathematics learning, but when and how to utilize this type of comparison merits additional research.

Aptitude–Treatment Interaction

The effectiveness of comparison varied with students' aptitude. Aptitude is not synonymous with intelligence or general ability; in the current study, condition did not interact with general math ability (as measured by a standardized math test) but rather with prior knowledge of solution methods. As noted by Snow (1992), aptitude is similar to readiness and encompasses a range of cognitive and affective characteristics of learners that influence how they respond to the particular demands and opportunities of a learning environment. Unfortunately, aptitude is a broad construct that does not provide specific guidance on which learner characteristics might be particularly important. Fortunately, prior research on the expertise-reversal effect provides more guidance, as it focuses on learners' experience with the target task as the appropriate measure of aptitude (Kalyuga, 2007). Indeed, our findings provide a new example of an expertise-reversal effect: the instructional approach that was most effective for novices in the domain was not most effective for more experienced learners. Comparison is another aspect of instruction that varies in effectiveness with learners' expertise in a domain, and learners' prior knowledge of target procedures is a key characteristic to attend to.

The What, When, and How of Comparison

Comparing solution methods and comparing problem types focused students' attention on different things, and prior knowledge of solution methods affected the two types of comparison differently. Clearly, what people are comparing matters, but very little attention in the literature is given to what dimensions differ between examples (Rittle-Johnson & Star, in press). When in the learning process people make comparisons matters, whether novices or experienced beginners, and this seems to impact learning differently depending on what is being compared. These issues of what and when have important implications for how comparison aids learning.

First, consider when and how comparison of solution methods might aid learning. This form of comparison may be particularly effective for learning a new method through analogy to a known method. According to theories of analogical learning, students can make inferences about a new example by identifying its similarities to and differences from a known example and making projec-

tions about how the new example works based on its alignment with the known example (Gentner, 1983; Hummel & Holyoak, 1997). Indeed, in past studies with more knowledgeable students, those who compared methods frequently identified how the unfamiliar, shortcut method was similar to and different from the distribute-first method that they already knew; in turn these types of comparative explanations predicted learning (Rittle-Johnson & Star, 2007, in press).

Few students in the current study knew the distribute-first method, so they could not use this type of analogy. Rather, they had to rely on mutual alignment of two unfamiliar methods (i.e., noticing potentially relevant features in two unfamiliar examples by identifying their similarities and differences; Gentner et al., 2003; Kurtz, Miao, & Gentner, 2001; Schwartz & Bransford, 1998). Even if students noticed important features of the examples, they would need sufficient knowledge to make sense of those features. For students who were unfamiliar with the general class of solution methods (i.e., algebraic methods rather than informal methods), this may simply have been too much. Solving equations is a complex process, involving multiple rules and variants; the processing load of learning the rules and variants simultaneously likely overwhelmed their working memory (Sweller, van Merrienboer, & Paas, 1998). Indeed, performance during the intervention suggested that comparing solution methods was overwhelming for students who did not use algebraic methods at pretest. These findings converge with prior research indicating that comparing two unfamiliar examples can be too difficult for young children (Gentner et al., 2007; Kotovsky & Gentner, 1996) or for college students who do not receive additional instructional support (Schwartz & Bransford, 1998). Before comparing multiple solution methods without teacher guidance, students may need to be familiar with one of the methods.

Next, consider when and how comparing problems might aid learning. For this type of comparison, students did seem to learn from mutual alignment of two unfamiliar examples. Perhaps it requires fewer resources to compare problem features than solution methods; comparing features of the equation requires only examining the first line of the example, whereas comparing methods requires comparing each line of the example and how the steps capitalize on problem features. Alternatively, the lack of an effect of prior knowledge on this type of comparison might be an artifact of our prior knowledge measure. Our measure focused on knowledge of solution methods, so it was more tailored to tapping prior knowledge differences particularly important in the comparing methods condition. An aptitude measure that captures familiarity with different types of equations might be more appropriate for evaluating Aptitude × Treatment interactions for comparing problems.

When and how different comparisons impact learning has important implications for theories of analogical learning. Analogy from a familiar to an unfamiliar example may aid learning more than does mutual alignment of unfamiliar examples, but the relative effectiveness of these two types of analogy has not been evaluated directly in prior research. Whether this is a general feature of analogical learning merits additional research. At the same time, theories of analogical learning must pay greater attention to what is being compared, as all comparisons are not created equal. They differ in their effectiveness and may depend on learners' prior knowledge and the complexity of the material.

Instructional Implications

The importance of prior knowledge for learning under different instructional conditions raises a critical instructional issue: For novices in a domain, what is the best way for students to learn new material? In the case of learning solution methods, one option would be for students to learn a new solution method by comparing it with a known, informal method. Indeed, one experimental curriculum focuses on bridging from arithmetic to algebra by having students develop and articulate their informal methods and then compare these methods with algebraic methods (Nathan, Stephens, Masarik, Alibali, & Koedinger, 2002). This instructional approach has facilitated students' ability to use and translate among multiple representations, relative to typical classroom instruction. We do caution that learning from this type of comparison will likely require extensive teacher support, given the difficulty of mapping between informal equation-solving methods and algebraic steps.

A second option is to have students learn a new solution method by comparing multiple examples of the same solution method used to solve different problems. Comparing highly similar examples prepares young and inexperienced learners to learn from less transparent comparisons later (Gentner et al., 2007; Kotovsky & Gentner, 1996) and may be particularly useful for novices in a domain. For example, students could compare parallel solutions to $3(x + 4) = 12$ and $7(y + 1) = 21$, which are easy to align and should help students learn the basic solution steps for a particular method.

A third option is for students to study examples of a single solution method individually before they do any comparison of examples. Novices in a domain often benefit from decomposition of complex tasks into smaller units (e.g., Lee & Anderson, 2001), presumably because decomposition avoids overloading working memory. This suggests that sequential presentation of individual examples may be best for not exceeding novices' capacity. In all three options, direct instruction and teacher-led whole-class discussion will likely facilitate learning for novices, as novices often benefit from high levels of instructional guidance (Cronbach & Snow, 1977; Kalyuga, 2007).

Future Directions and Conclusion

Several lines of research would help confirm and explicate the importance of prior knowledge for learning from comparison. First, to directly confirm the importance of prior knowledge of a solution method, novices in a domain should be randomly assigned to learn and practice one solution method before comparing the method with a second method or to compare methods from the beginning. All three of the methods for learning a single method described above—comparison of a new method with a known informal method, comparison of multiple examples of the same solution method used to solve different problems, and sequential study and practice of one method—should be evaluated to assess whether different types of comparison can be used to build competence with a single method. Any method that supports competence with one solution method should prepare students to learn from comparing solution methods, but different methods might be more efficient or might support better generalization. Second, we need to know how much prior knowledge is sufficient and how to

measure this prior knowledge. For example, Kalyuga and Sweller (2004) developed rapid knowledge assessments that were successfully used to identify whether middle-school students should learn about coordinate geometry or equation solving by studying worked examples or by solving problems. Their rapid knowledge assessments asked students to identify the next step in a problem solution and gave higher scores to students who skipped intermediate steps. Our own past work suggests that successful use of one of the target solution methods on at least one problem is a good indicator that students are ready to compare the method to a new method. Both lines of work suggest that students' solution methods are a good indicator of prior knowledge. Third, the effectiveness of different types of comparison when additional instructional support is provided, such as teacher-led whole class discussions, is greatly needed. Novices in a domain typically need greater structure, completeness of information, and direct teaching methods than do more knowledgeable learners (Cronbach & Snow, 1977). Such methods should be tried with comparison. Finally, the generalizability of these findings, both to other domains and to more typical classroom conditions, is greatly needed.

In conclusion, it matters what and when students compare. For those with little prior knowledge of solution methods, it is best to delay comparisons of multiple solution methods. Rather, students should become familiar with at least one method, either through sequential study of examples illustrating the method or through comparison of different problems solved with the same solution method. Once students can use one method, comparing that method to an alternative method should improve their procedural flexibility and conceptual knowledge (Rittle-Johnson & Star, 2007, in press; Star & Rittle-Johnson, 2009).

References

Albro, E., Uttal, D., De Loache, J., Kaminski, J. A., Sloutsky, V. M., Heckler, A. F., et al. (2007). Fostering transfer of knowledge in education settings. In D. S. McNamara & G. Trafton (Eds.), *Proceedings of the 29th meeting of the Cognitive Science Society* (pp. 21–22). Austin, TX: Cognitive Science Society.

Ball, D. L. (1993). With an eye on the mathematical horizon: Dilemmas of teaching elementary school mathematics. *Elementary School Journal, 93,* 373–397.

Barzi, F., & Woodward, M. (2004). Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology, 160,* 34–45.

Beishuizen, M., van Putten, C. M., & van Mulken, F. (1997). Mental arithmetic and strategy use with indirect number problems up to one hundred. *Learning and Instruction, 7,* 87–106.

Bjorklund, D. F., Miller, P. H., Coyle, T. R., & Slawinski, J. L. (1997). Instructing children to use memory strategies: Evidence of utilization deficiencies in memory training studies. *Developmental Review, 17,* 411–441.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals.* New York: McKay.

Blöte, A. W., Van der Burg, E., & Klein, A. S. (2001). Students' flexibility in solving two-digit addition and subtraction problems: Instruction effects. *Journal of Educational Psychology, 93,* 627–638.

Bransford, J. D., Franks, J. J., Vye, N. J., & Sherwood, R. D. (1989). New approaches to instruction: Because wisdom can't be told. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 470–497). New York: Cambridge University Press.

Carpenter, T. P., Franke, M. L., Jacobs, V. R., Fennema, E., & Empson, S. B. (1998). A longitudinal study of invention and understanding in children's multidigit addition and subtraction. *Journal for Research in Mathematics Education, 29,* 3–20.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 1147–1156.

Clarke, T., Ayres, P., & Sweller, J. (2005). The impact of sequencing and prior knowledge on learning mathematics through spreadsheet applications. *Educational Technology Research and Development, 53*(3), 15–24.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

Cummins, D. (1992). Role of analogical reasoning in the induction of problem categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1103–1124.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170.

Gentner, D. (2005). The development of relational category knowledge. In D. H. Rakison & L. Gershkoff-Stowe (Eds.), *Building object categories in developmental time* (pp. 245–275). Mahwah, NJ: Erlbaum.

Gentner, D., Loewenstein, J., & Hung, B. (2007). Comparison facilitates children's learning of names for parts. *Journal of Cognition and Development, 8,* 285–307.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95,* 393–405.

Gentner, D., & Namy, L. L. (2004). The role of comparison in children's early word learning. In S. R. Waxman & D. G. Hall (Eds.), *Weaving a lexicon* (pp. 533–568). Cambridge, MA: MIT Press.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1–38.

Hiebert, J., & Wearne, D. (1996). Instruction, understanding, and skill in multidigit addition and subtraction. *Cognition and Instruction, 14,* 251–283.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104,* 427–466.

Johnson, D. W., & Johnson, R. T. (1994). *Learning together and alone: Cooperative, competitive and individualistic learning* (4th ed.). Boston: Allyn & Bacon.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19,* 509–539.

Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. *Journal of Educational Psychology, 96,* 558–568.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis.* New York: Guilford Press.

Kieran, C. (1992). The learning and teaching of school algebra. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 390–419). New York: Simon & Schuster.

Kilpatrick, J., Swafford, J. O., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics.* Washington, DC: National Academy Press.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67,* 2797–2822.

Kurtz, K., Miao, C.-H., & Gentner, D. (2001). Learning by analogical bootstrapping. *Journal of the Learning Sciences, 10,* 417–446.

Lampert, M. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. *American Educational Research Journal, 27,* 29–63.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D.

(2009). *Connected mathematics 2.* Upper Saddle River, NJ: Pearson Education.

Lee, F. J., & Anderson, J. R. (2001). Does learning a complex task have to be complex? A study in learning decomposition. *Cognitive Psychology, 42,* 267–316.

Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General, 131,* 5–15.

Nathan, M. J., Stephens, A. C., Masarik, K., Alibali, M. W., & Koedinger, K. R. (2002). Representational fluency in middle school: A classroom study. In D. Mewborn, P. Sztajn, D. White, H. Wiegel, R. Bryant, & K. Nooney (Eds.), *Proceedings of the 24th annual meeting of the North American Chapter for the Psychology of Mathematics Education* (pp. 463–472). Columbus, OH: ERIC Clearinghouse for Science, Mathematics and Environmental Education.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through Grade 8 mathematics.* Reston, VA: Author.

National Mathematics Advisory Panel. (2008). *Foundations of success: The final report of the National Mathematics Advisory Panel.* Washington, DC: U.S. Department of Education.

Oakes, L. M., & Ribar, R. J. (2005). A comparison of infants' categorization in paired and successive presentation familiarization tasks. *Infancy, 7,* 85–98.

Paas, F. G. W. C., & Van Merrienboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86,* 122–133.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., et al. (2005). The assistment project: Blending assessment and assisting. In C. K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence in Education* (pp. 555–562). Amsterdam: ISO Press.

Richland, L. E., Holyoak, K. J., & Stigler, J. W. (2004). Analogy use in eighth-grade mathematics classrooms. *Cognition and Instruction, 22,* 37–60.

Richland, L. E., Zur, O., & Holyoak, K. J. (2007, May 25). Cognitive supports for analogies in the mathematics classroom. *Science, 316,* 1128–1129.

Rittle-Johnson, B., & Kmicikewycz, A. O. (2008). When generating answers benefits arithmetic skill: The importance of prior knowledge. *Journal of Experimental Child Psychology, 101,* 75–81.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93,* 346–362.

Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology, 99,* 561–574.

Rittle-Johnson, B., & Star, J. R. (2009). Compared to what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. *Journal of Educational Psychology, 101,* 529–544.

Satterthwaite, F. E. (1946). An approximate distribution of estimation of variance components. *Biometrics Bulletin, 2,* 110–114.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16,* 475–522.

Silver, E. A., Ghousseini, H., Gosen, D., Charalambous, C., & Strawhun, B. (2005). Moving from rhetoric to praxis: Issues faced by teachers in having students consider multiple solutions for problems in the mathematics classroom. *Journal of Mathematical Behavior, 24,* 287–301.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill.* Cambridge, MA: Harvard University Press.

Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational Psychologist, 27,* 5–32.

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education, 36,* 404–411.

Star, J. R., & Rittle-Johnson, B. (2008). Flexibility in problem solving: The case of equation solving. *Learning and Instruction, 18,* 565–579.

Star, J. R., & Rittle-Johnson, B. (2009). It pays to compare: An experimental study on computational estimation. *Journal of Experimental Child Psychology, 101,* 408–426.

Star, J. R., & Seifert, C. (2006). The development of flexibility in equation solving. *Contemporary Educational Psychology, 31,* 280–300.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10,* 251–296.

VanderStoep, S. W., & Seifert, C. M. (1993). Learning "how" versus learning "when": Improving transfer of problem-solving principles. *Journal of the Learning Sciences, 3,* 93–111.

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education, 22,* 366–389.

Received December 6, 2008
Revision received March 30, 2009
Accepted March 31, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 853–866

0022-0663/09/$12.00 DOI: 10.1037/a0016306

Within-School Social Comparison: How Students Perceive the Standing of Their Class Predicts Academic Self-Concept

Ulrich Trautwein and Oliver Lüdtke
University of Tuebingen and Max Planck Institute for Human Development

Herbert W. Marsh
Oxford University

Gabriel Nagy
Max Planck Institute for Human Development

Results from prior research indicate that a student's academic self-concept is negatively influenced by the achievement of others in his or her school (a frame of reference effect) and that this negative frame of reference effect is not or only slightly reduced by the quality, standing, or prestige of the track or school attended (a "reflected glory" effect). Going beyond prior studies, the present research used both between-school and within-school approaches to investigate frame of reference and reflected glory effects in education, incorporating students' own perceptions of the standing of their school and class. Multilevel analyses were performed with data from 3 large-scale assessments with 4,810, 1,502, and 4,247 students, respectively. Findings from all 3 studies showed that, given comparable individual achievement, placement in high-achieving learning groups was associated with comparatively low academic self-concepts. However, students' academic self-concept was not merely a reflection of their relative position within the class but also substantively associated with their individual and shared perceptions of the class's standing. Moreover, the negative effects of being placed in high-achieving learning groups were weaker for high-achieving students. Overall, the studies support both educational and social psychology theorizing on social comparison.

Keywords: social comparison, achievement, frame of reference effect, reflected glory effect

A high evaluation of one's skills and abilities in important academic and nonacademic life domains contributes to high self-esteem and life satisfaction (Taylor & Brown, 1988; Trautwein, Lüdtke, Köller, & Baumert, 2006). Furthermore, feeling competent in a specific area motivates and energizes behavior in that domain and is associated with favorable long-term outcomes (Bandura, 1997; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005; Trautwein, Lüdtke, Kastens, & Köller, 2006). Not surprisingly, the sources of such positive self-evaluations have been the subject of much research. Several studies have shown that the immediate environment constitutes a salient frame of reference that impacts people's self-evaluations (Suls & Wheeler, 2000). In our own research (e.g., Marsh, 1987; Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006), we have studied frame of reference effects primarily within educational environments, where social comparison processes are widespread. Frame of reference effects in educational environments have considerable implications for students' lives, affecting outcomes such as their long-term educational trajectories (Marsh, 1991; Trautwein & Baeriswyl, 2007) and health-related behaviors (Trautwein, Gerlach, & Lüdtke, 2008). Because the complex characteristics of educational environments involving many students and multiple frames of reference pose a challenge for social comparison theories focusing on individual or dyadic processes, moreover, research on frame of reference effects in natural school environments makes an important contribution to the literature on social comparison processes.

In this article, we examine the possible impact of frame of reference effects in secondary school. Secondary schools are known to differ markedly in their student composition and overall achievement (Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007; Trautwein, Lüdtke, Marsh, et al., 2006)—factors that are likely to impact students' self-evaluations. Results from our prior research and from several other studies (e.g., Tymms, 2001; Zeidner & Schleyer, 1999) indicate that a student's academic self-concept is strongly influenced by the achievement of others in his or her school and that this frame of reference effect also applies to students' interest, course choice, and educational aspirations (e.g., Marsh, 1991; Trautwein, Köller, Lüdtke, & Baumert, 2005; Trautwein, Lüdtke, Marsh, et al., 2006). However, the social comparison processes underlying this frame of reference effect are not yet fully understood (Wheeler & Suls, 2007). The present article makes a fourfold contribution to the literature. First, whereas most previous research has used between-school designs to investigate frame of reference effects in education, we also address within-school differences (i.e., differences between classes in the same school). Second, in addition to objective information

Ulrich Trautwein and Oliver Lüdtke, Department of Education, University of Tuebingen, Tuebingen, Germany, and Center for Educational Research, Max Planck Institute for Human Development, Berlin, Germany; Herbert W. Marsh, Department of Education, Oxford University, Oxford, England; Gabriel Nagy, Center for Educational Research, Max Planck Institute for Human Development.

Correspondence concerning this article should be addressed to Ulrich Trautwein, University of Tuebingen, Muenzgasse 22-26, 72070 Tuebingen, Germany. E-mail: ulrich.trautwein@uni-tuebingen.de

about the relative standing of a class or school, we investigate students' individual and shared perceptions of their class or school's standing. To what extent do students agree about the overall standing of their class or school, and how are these perceptions related to objective achievement? Third, we relate these perceptions of the standing of the class or school to students' academic self-concepts, testing whether a positive perception of class or school quality counteracts the expected negative frame of reference effect. Fourth, we probe for possible interaction effects between the average achievement of a class or school and the perceived standing of the class or school, on the one hand, and individual achievement, on the other. In other words, we test whether high-achieving and low-achieving students are differentially affected by class characteristics.

Research Paradigms in Social Comparison Research

Psychological research on social comparison dates back to James (1890/1963) and Festinger (1954). Disciplines such as social, developmental, and educational psychology have since seen a rich history of social comparison research; each discipline has focused on different phenomena and used different research designs (see Suls & Wheeler, 2000; Wheeler & Suls, 2005; Wood & Wilson, 2003). In social psychology, there has been an emphasis on experimental studies on the effects of upward and downward social comparison, the motives for these comparison processes, and the preference for and differential impact of specific types of comparison information. The focus is on the individual as a processor of social comparison information. Typically, the respondent is either presented with a specific target or asked to pick a target. In either case, the interest is in individual characteristics rather than in group characteristics.

Results from this body of research suggest that people are constantly on the lookout for social comparison information that can be integrated into their self-concepts (Suls & Wheeler, 2000; for a critical review, see Wood & Wilson, 2003). Perhaps the most intriguing finding is the compelling evidence for self-enhancing mechanisms in social comparison processes (e.g., Damisch, Mussweiler, & Plessner, 2006; Mussweiler, 2003; Suls & Wheeler, 2000). Research has shown that—under certain conditions—social comparison with a target who shows greater proficiency in a specific domain may result in the comparer developing a higher self-evaluation through assimilation. Assimilative processes are especially likely if the comparer and the target share important characteristics (Mussweiler, 2003; Wheeler & Suls, 2005). In somewhat simplified terms, social psychology research indicates that humans are capable of using social comparison processes adaptively to enhance their self-evaluations.

Research on social comparison processes in educational psychology departs from the paradigms used in social psychology (e.g., Wheeler & Suls, 2005). One major strand of research has focused on the impact of achievement differences in naturally occurring educational environments (e.g., classes, schools) on outcome variables such as academic self-concept and educational choices (Marsh & Craven, 2002). In this paradigm, student outcomes are seen as the consequence of specific characteristics of the (natural) learning environment. Unlike laboratory experiments in social psychology, where the experimenter manipulates the available social comparison information, real-life educational settings provide a wealth of potentially useful social comparison information, and researchers seek to identify the most important sources of information by relating characteristics of the learning environment to student outcomes.

An important feature of typical educational settings is their hierarchical structure: Students are nested within classes, classes are nested within schools, and schools are nested within larger units such as school districts, states, or countries. It is imperative to distinguish between these hierarchical levels for both conceptual and statistical reasons (Raudenbush & Bryk, 2002). Although researchers on social comparison processes in educational psychology are well aware of these different levels of analyses, not all levels have been covered in similar detail, as we describe in more detail below.

The social comparison paradigm used in educational psychology has yielded a fairly consistent body of results (Marsh & Craven, 2002). However, in contrast to the overall picture emerging from social psychology research—in which high-achieving comparison targets often activate assimilation processes with positive effects on the comparer's self-concept (see Mussweiler, 2003)—the vast majority of studies in natural learning environments have found that high-achieving schoolmates have negative effects on their fellow students' self-concepts. In the following, we describe some of the studies and findings most relevant to the present research.

Imposed Social Comparison: Frame of Reference Effects in the Classroom

The analytic approach most frequently chosen in studies examining frame of reference effects in educational environments is regression based. General or domain-specific academic self-concept (assessed by items such as "I am smart" or "I am good in mathematics") is used as the outcome variable, and individual student achievement and school-average achievement are used as the two major predictor variables. Regression analysis is used to test whether school-average achievement is positively or negatively associated with self-concept when individual achievement is statistically controlled. In other words, it examines the consequences of placement in high- or low-achieving environments. Given two students with comparable achievement scores, which student has a higher academic self-concept: the one placed in a high-achieving school or the one placed in a low-achieving school? The large majority of studies have found a negative regression coefficient of school-average achievement as measured by standardized achievement tests on academic self-concept (e.g., Lüdtke, Köller, Marsh, & Trautwein, 2005; Marsh & Hau, 2003; Marsh, Köller, & Baumert, 2001; see also the review by Marsh & Craven, 2002), a phenomenon known as the "big-fish-little-pond effect" (BFLPE; see Marsh, 1987, 1991; Marsh & Hau, 2003).

The empirical support for the BFLPE is compelling. For instance, Marsh and Hau (2003) conducted a large cross-cultural test of frame of reference effects using data from the Programme for International Student Assessment (PISA; Organization for Economic Cooperation and Development, 2001). Nationally representative samples of approximately 4,000 students from each of the 26 participating countries (total N = 103,558 students in 3,851 schools) completed standardized achievement tests and a self-concept questionnaire. Consistent with a priori predictions, the

predictive effects of individual student achievement were substantial and positive, whereas the regression coefficients for school-average achievement were negative.

Some researchers have used track status rather than school-average achievement to predict self-concept. In a German study, Schwarzer, Lange, and Jerusalem (1982) examined the effect of track status on the development of academic self-concept after transition to secondary school. Students in Germany are tracked on the basis of their achievement at about age 10. The academic self-concept of high-achieving students (who were placed in the high track) tended to decrease after transition to secondary school, whereas the academic self-concept of low-achieving students (who were placed in the low track) tended to increase, indicating that the negative effect of high-achieving classmates was stronger than any positive effect of high track membership. Similarly, Rheinberg and Enstrup (1977) compared the academic self-concept, test anxiety, and achievement motivation of 165 students with mild to moderate learning disabilities ($70 < IQ \leq 85$). When achievement was controlled, students attending special schools were found to have higher academic self-concepts and achievement motivation and lower test anxiety than those enrolled in regular schools.

Counterbalancing Effects: Does the Standing of the School Predict Self-Concept?

The studies reported thus far indicate that the self-concept of students who are placed in academically selective schools is negatively affected—a negative BFLPE or contrast effect. However, might self-perceptions not also be enhanced by membership of high-achieving or positively valued groups? In the social psychology literature, there is sound evidence that people enjoy basking in the reflected glory of successful others (e.g., Cialdini & Richardson, 1980) and that self-perceptions may be enhanced by membership in groups that are positively valued by the individual (Diener & Fujita, 1997; Tesser, 1988). Adopting the term *reflected glory effects*, Marsh (1984, 1987; Marsh, Kong, & Hau, 2000) argued that—theoretically speaking—students in academically selective schools might have more positive academic self-concepts by virtue of being affiliated with a highly selective educational program. In this sense, placement in a high-achievement group might be expected to positively affect students' global and domain-specific self-concepts by means of "assimilation effects" (see Marsh et al., 2000; Oakes, 1985; Seaton et al., 2008). From the theoretical point of view, these reflected glory effects might weaken or fully counterbalance negative frame of reference effects.

There are three major approaches to testing the relative strength of reflected glory and negative frame of reference effects. The first approach is used when no information other than average school achievement is available on relevant school characteristics. In this case—i.e., in the majority of studies on the BFLPE—reflected glory and negative frame of reference effects are confounded in the regression coefficient of school-average achievement. Because the total effect (the regression coefficient of school-average achievement) is almost always negative (see Marsh, Seaton, et al., 2008), it is evident that the negative effect is stronger than the positive effect, although the size of each effect is unknown.

The second approach is used when additional descriptive information about a school or a class is available. The best example is information on within-school or between-school tracking. In tracked school systems, students are assigned to a specific track on the basis of their prior achievement, leading to homogenization of learning groups (see Maaz, Trautwein, Lüdtke, & Baumert, 2008). Students in higher tracks benefit from more cognitively activating instruction and are more likely to gain access to university (Becker, Lüdtke, Trautwein, Köller, & Baumert, 2008; Klusmann, Kunter, Trautwein, Lüdtke, & Baumert, 2008; Maaz et al., 2008). Hence, membership in a higher track may produce reflected glory effects. School-average achievement and track level are highly, but not perfectly, correlated. Accordingly, when track information is included in the analyses, in addition to school- or class-average achievement, regression analyses should separate the (negative) frame of reference effects (as mirrored in school-average achievement) from the (positive) reflected-glory effects (as expressed in track status). To date, only a handful of studies have used this approach. Recently, Trautwein, Lüdtke, Marsh, et al. (2006, Study 1) included both track membership and school-average achievement as predictor variables in a study with more than 14,000 ninth graders in Germany. When individual achievement and students' teacher-assigned school grades were controlled, school-average achievement negatively predicted academic self-concept to a statistically significant degree, whereas membership in a high or low track was not associated with self-concept. The authors interpreted this finding as indicating that students typically integrate information about the achievement of other students in their school into their self-concept, but not information about the achievement of students in other schools or the prestige of their track.

The third approach to separating the negative and positive effects of being placed in a selective learning environment uses student perceptions of the standing of their school or class as an additional predictor variable. These perceptions can additionally be aggregated to the class or school level and correlated with objective achievement and self-concept. This approach thus draws on both the individual perspective and the shared perceptions of a group of students. Accordingly, from a psychological point of view, it has the greatest potential for modeling reflected glory effects. To our knowledge, however, only a single published study has used this approach. Marsh et al. (2000) followed a large, nationally representative sample of Grade 7 students through high school in Hong Kong (7,997 students, 44 high schools, 4 years). Although Hong Kong does not have a classical tracked school system, parents and students are well aware of each school's relative standing, and they use this information when selecting schools. The availability of this information might enable students in selective high schools to maintain a favorable self-concept despite their constant exposure to high-achieving fellow students. Indeed, as expected by the authors, the higher the school-average achievement, the higher the perceived school status reported by the students. Consistent with previous findings of negative frame of reference effects, when individual achievement was controlled, school-average achievement based on measures collected in Grade 6, prior to the transfer to high school, negatively predicted academic self-concept in Grade 8 and Grade 9. Most important in the present context, however, individual students' perceptions of the status of their school positively predicted their academic self-concept, counterbalancing some of the negative effects of being placed in a selective environment.

The Marsh et al. (2000) study indicates that students' academic self-concepts are not fully determined by their relative position in school, but also reflect their beliefs about the relative standing of their school. Unfortunately, however, the Marsh et al. study did not examine whether students within a school had similar perceptions of the school's status or whether their perceptions were idiosyncratic. Moreover, the study did not analytically separate the effects of individual (idiosyncratic) perceptions, on the one hand, and perceptions shared by the students within a school, on the other. This distinction is of high theoretical and empirical interest. If students with higher perceptions of their school's status have higher self-concepts, the school status effect reported reflects an individual-level effect. Alternatively, if the mean academic self-concept is higher in schools with a relatively high mean perception of school status, it reflects a school-level effect. Because the Marsh et al. study did not analytically distinguish between the individual and school levels, there is no way of telling whether the school status effect they found documented an effect at the individual level, the school level, or a mixture of both (Cronbach, 1976; Lüdtke, Robitzsch, Trautwein, & Kunter, 2009).

Do Class Characteristics Interact with Individual Student Characteristics?

Another important issue in the study of reference group effects is whether these effects apply to all students in the same way. In other words, if—when individual ability is controlled—there is a negative regression coefficient of school/class-average ability on student self-concept, is this effect the same for all students within a class? Are high- and low-achieving students within a class equally affected by a high average ability of their reference group?

From a theoretical perspective, Marsh and colleagues (Marsh, 1987; 1991; Marsh, Trautwein, Lüdtke, & Köller, 2008) argued that interactions between school/class-average ability and individual ability on academic self-concept might be relatively small or nonsignificant because the frame of reference is established by school/class-average ability. Accordingly, all students in a high-ability school/class are predicted to have lower academic self-concepts than they would if they attended a low-ability school/class.

From the empirical perspective, the relatively few studies investigating whether the BFLPE is similar at all ability levels have yielded nonsignificant results or relatively small effects. Moreover, not even the direction of the small effects was consistent across studies. For instance, the findings of Marsh, Chessor, Craven, and Roche (1995) and Marsh and Hau (2003) suggest that the BFLPE affects all levels of ability in a similar way. Marsh et al. (2007) tested interaction effects between school-average ability and individual ability in two samples of college-track high school students. Whereas there was no evidence for an interaction effect in the first sample, a negative interaction term in the second sample suggested that high-achieving students were more strongly affected by placement in high-achieving schools. Conversely, Dai and Rinn (2008) argued that findings from some studies of gifted student programs indicate that students in these selective academic programs may be less affected by negative frame of reference effects. In sum, although results to date are inconclusive (Coleman & Fults, 1985; Dai & Rinn, 2008; Marsh et al., 2007; Reuman, 1989), individual difference in ability is a potentially important BFLPE moderator that warrants further consideration.

Apart from its achievement level, the perceived standing of a class or school may also interact with individual student achievement. Believing oneself to be part of a high-quality learning group may be differentially important for high-or low-achieving students. For instance, it is conceivable that the self-concept of a low-achieving student benefits more from the idea of belonging to a high-quality class or school than does the self-concept of a high-achieving student. However, empirical studies have yet to test this hypothesis.

The Present Research

The majority of previous studies on frame of reference effects in educational settings support the hypothesis that when individual achievement is controlled, placement in a high-achievement group is associated with lower academic self-concepts. However, only one study to date has examined whether student ratings of their school's standing influence their self-perceptions. This one exception, the pioneering study by Marsh et al. (2000), considered student ratings of school status but did not separate individual- and school-level effects. Thus, previous research on self-concept in educational environments strongly supports the hypothesis that the school is the most salient frame of reference, but very little attention has been paid to student perceptions of the school or the track. This paucity of empirical studies stands in marked contrast to the picture that has emerged from social psychology, where the bulk of laboratory research portrays humans as active information seekers who adaptively use the social comparison information available (Mussweiler, 2003; Suls & Wheeler, 2000) to enhance their self-concept.

The present research builds on prior research to critically assess social comparison processes in natural learning environments. Most important, we included a measure of perceived standing of the class or school in our analyses, asking students to evaluate the standing of their mathematics class relative to other mathematics classes in the school (Studies 1 and 2) and the standing of their school relative to other schools (Study 3).

In principle, we tested the following set of four questions in all three studies (in Study 3, we looked at school characteristics rather than class characteristics): First, we expected to replicate results from prior research on frame of reference effects, independently of whether class or school was used as the grouping variable. In other words, when individual mathematics achievement was controlled, we expected to find a negative regression weight of class- or school-average mathematics achievement on mathematics self-concept. Second, extending the Marsh et al. (2000) study, we distinguished between students' individual perceptions of the standing of their class or school and their shared perceptions of the standing of their class or school. We expected to find reliable between-class and between-school differences in students' evaluations of the standing of their class and school. That is, we expected students in the same classes or schools to report some shared perceptions of their class or school's standing, although individual members of the same class or school were naturally expected differ to some degree in their evaluations. Third, we expected these ratings of the standing of the class or school to be reflected in students' mathematics self-concepts, and we made

parallel predictions for the class or school and the individual levels. At the class or school level, we expected to find higher mathematics self-concept in classes in which the overall student rating of class or school standing was high. Similarly, at the student level, we expected to find higher mathematics self-concept in students who reported higher perceptions of their class or school's standing. Finally, we probed for possible interaction effects between two class or school characteristics (the average achievement of a class or school and the perceived standing of the class or school), on the one hand, and individual student characteristics (student achievement and perceived class or school standing), on the other. Hence, four interaction effects in total were specified in each of the three studies. We were specifically interested in the interaction effect between class- or school-average achievement and individual achievement because—as described above—prior research has yielded inconclusive results.

Study 1

Method

Sample

The data for Study 1 were collected in 156 randomly selected academic-track secondary schools in the German state of Baden-Württemberg that were representative of the track in that state. Students graduating from academic-track secondary schools in Germany are eligible to attend university. Forty students in each school were randomly selected and invited to participate in the study; in schools with less than 40 students, all students were invited to participate. Of the targeted sample of 6,177 students, a total of $N = 5{,}016$ students (= 81.2%) participated in the study. Complete data were available from 4,810 students (56% female; mean age $M = 19.57$ years, $SD = 0.78$); these students form the sample for the present study. Students in the present sample were in their final year of schooling (i.e., Grade 13); all were enrolled in a pre-university mathematics class. Students participated voluntarily in the present investigation without any financial reward. Two trained research assistants administered materials in each school between February and May 2006.

Instruments

Mathematics self-concept. Mathematics self-concept was assessed using the German adaptation of the Self Description Questionnaire III, a multidimensional self-concept instrument for late adolescents and young adults based on the Shavelson, Hubner, and Stanton (1976; Marsh & Shavelson, 1985) model. In the German adaptation (Schwanzer, Trautwein, Lüdtke, & Sydow, 2005), four researchers with English as a second language translated all original items independently of each other. Subsequently, the most appropriate translation was chosen (and in some instances refined) with the assistance of a professional translator. Extensive pilot testing resulted in a short German instrument with 4 items per scale and a 4-point response format (from *disagree* to *agree*). The 4 items selected per scale emphasized cognitive (e.g., "I'm good at mathematics") rather than affective (e.g., "I like mathematics") evaluations. Marsh et al. (2007) have provided strong empirical support for the convergent and discriminant validity of responses to this mathematics self-concept scale. In the present study, the internal consistency (Cronbach's alpha) of the scale score was .90.

Mathematics achievement. The mathematics achievement test consisted of items from the Third International Mathematics and Science Study (TIMSS; e.g., Baumert, Bos, & Lehmann, 2000). Responses were analyzed with item response theory methods and the ConQuest software (Wu, Adams, & Wilson, 1998), and the original TIMSS metric was used to generate a total mathematics achievement score. The reliability of the test scores was .88 (formula by Rost, 1996).

Perceived standing of the mathematics class. Four statements with an identical item stem ("Relative to other mathematics classes in my school ___") tapped students' ratings of the standing of their mathematics class ("my mathematics class is considered to be good," "we learn a lot in my mathematics class," "students in my mathematics class are particularly high achieving," "my mathematics class has high standing").[1] A 4-point (from *disagree* to *agree*) response format was used. The internal consistency (Cronbach's alpha) of the scale score was .85. The measure was partly adapted from the Marsh et al. (2000) study. An exploratory factor analysis strongly supported a one-factor solution (all factor loadings on the first component were > .73).

Statistical Analyses

In most studies conducted in schools, individual student characteristics are confounded with classroom or school characteristics because individual students are not randomly assigned to groups. This clustering effect introduces problems related to appropriate levels of analysis, aggregation bias, and heterogeneity of regression (Raudenbush & Bryk, 2002). We therefore performed multilevel regression analyses to predict mathematics self-concept. For the present investigation, it is particularly important to note that the meaning of a variable at the student level may not bear any straightforward relation to its meaning at the classroom level. The negative frame of reference effect represents a dramatic example of this problem, in that achievement at the individual level is positively related to self-concept, whereas achievement at the class-average level may be unrelated or negatively related to self-concept. The juxtaposition of the effects of individual achievement and class-average achievement is inherently a multilevel issue that cannot be represented properly at either the individual or the classroom level. Particularly when major variables represent different levels, it is important to use appropriate multilevel statistical procedures for data analysis. Multilevel modeling, a general form of regression analysis, provides a powerful methodology for handling hierarchical data, and it was used in this study. A detailed presentation of multilevel modeling (also referred to as hierarchical linear modeling [HLM]) is beyond the scope of the present investigation and is available elsewhere (e.g., Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).

[1] The original German wording of the item stem and the four statements (items) was as follows: Item stem: "*Im Vergleich mit anderen Mathematikkursen an meiner Schule* ___"; items: "*hat mein Mathematikkurs einen guten Ruf*," "*wird in meinem Mathematikkurs viel gelernt*," "*sind Schülerinnen und Schüler in meinem Mathematikkurs besonders leistungsstark*," "*ist mein Mathematikkurs besonders angesehen*."

In the present study, all multilevel analyses were computed using the HLM 6 computer program (Raudenbush, Bryk, Cheong, & Congdon, 2004). We specified three-level models, with students as the first level, classes as the second level, and schools as the third level. We did not, however, include school-average achievement as an additional predictor because the very high intercorrelation of $r = .80$ between class-average achievement and school-average achievement would produce unwanted multicollinearity if both variables were introduced simultaneously, making results very difficult to interpret.

We specified several sets of multilevel models to test our hypotheses. As in ordinary regression analyses, one outcome variable was regressed on several predictor variables in each model. By specifying several consecutive models, researchers are able to observe the change in the predictive power of one variable when an additional variable is included. HLM does not report standardized regression coefficients. To enhance the interpretability of the regression coefficients produced, we standardized ($M = 0$, $SD = 1$) all continuous variables before performing the multilevel analyses. Mathematics achievement was aggregated at the class level (Level 2) to form an index of the overall level of mathematics achievement in the class (and was not restandardized). Unless otherwise indicated, all models reported are random-intercept models estimated by the full maximum likelihood method.

Centering Level 1 (student-level) predictor variables is a crucial issue in multilevel modeling (see Enders & Tofighi, 2007). In the literature on multilevel modeling, two main centering options are discussed: Level 1 predictor variables can be adjusted to the mean of the cluster to which the student belongs (centering at the group mean) or to the mean of the variables in the whole sample (centering at the grand mean). In line with prior research on frame of reference effects, students' individual achievement was centered at the grand mean. Thus, the predictive effect of class-average mathematics achievement on mathematics self-concept is controlled for individual mathematics achievement. In contrast, perceived class standing was centered at the group mean. Centering at the group mean is typically the appropriate option for students' ratings of their learning environment (Lüdtke et al., 2009): Grand mean centering would lead to interindividual differences among classes being controlled in these ratings, thereby eliminating an essential component of the aggregated ratings (see also Karabenick, 2004).

We assessed model fit using the deviance values provided by HLM, which can be regarded as a measure of lack of fit between model and data (Snijders & Bosker, 1999). Deviance values are not usually interpreted directly; rather, differences in deviance values are calculated for models applied to the same data set. The difference in deviance between two models has a chi-square distribution with degrees of freedom equal to the difference in the number of parameters estimated. Because we used the full maximum likelihood method, the chi-square statistic can be used to evaluate change in model fit when either a fixed or a random effect is added. Large chi-square values indicate that the model with more parameters provides a better fit to the data than the more parsimonious model.

Effect sizes have found increasing use in psychological research (see Grissom & Kim, 2005). In our study, we used three indicators of effect size. First, in analogy to the measure of explained variance in ordinary linear regression models, we report the overall proportion of variance explained by the predictor variables for each model. This measure is determined by calculating the decrease in the total variance when the predictor variables are introduced into the specific model (see Snijders & Bosker, 1999).

Second, we report easily interpretable regression coefficients for Level 1 variables. Because we standardized all continuous Level 1 predictor and outcome variables before entering them in our multilevel models, the coefficients of the continuous Level 1 variables can be interpreted in almost the same way as the standardized regression coefficients resulting from ordinary regression analysis. With reference to Cohen's (1988) suggestions for correlations, we consider a regression coefficient of .10 to mark the lower bound for a meaningful effect (for a similar rationale, see Roberts, Caspi, & Moffitt, 2003).

Third, in line with most previous multilevel research (e.g., Marsh & Hau, 2003), we did not restandardize our Level 2 predictor variable (class-average mathematics achievement). Hence, the class-level regression weights show change in the dependent variable (mathematics self-concept) corresponding to an increase of one unit in class-average mathematics achievement, expressed in the metric of mathematics achievement at the student level. Tymms (2004) proposed that the effect size for continuous Level 2 predictors in multilevel models, which is comparable with Cohen's d, be calculated using the following formula:

$$\Delta = 2 \times B \times SD_{\text{predictor}}/\sigma_e,$$

where B is the unstandardized regression coefficient in the multilevel model, $SD_{\text{predictor}}$ is the standard deviation of the predictor variable at the class level, and σ_e is the residual standard deviation at the student level. The resulting effect size describes the difference in the dependent variable between two classes that differ by two standard deviations on the predictor variable. We suggest that an effect size of $\Delta \geq |0.20|$ can be considered of practical significance in the present research.

Results and Discussion

Our first hypothesis stated that the typical frame of reference effect on academic self-concept would also be found when the class—rather than the school—was used as the grouping variable. In other words, we expected to find a negative regression weight for class-average achievement when individual achievement was controlled. In Model 1, we therefore specified the classical frame of reference model, including both individual mathematics scores and class-average mathematics scores as predictor variables. As documented in Table 1, individual mathematics achievement positively predicted mathematics self-concept; the regression coefficient of $B = 0.60$ indicates that an increase in mathematics achievement of one standard deviation was associated with an increase in mathematics self-concept of more than half a standard deviation. The regression coefficient for class-average mathematics achievement was $B = -0.22$. We calculated the effect size for this coefficient using the formula given above. With a standard deviation in class-average ability of $SD = 0.59$ and a residual standard deviation at the student level of $\sigma_e = 0.83$, we found a small effect:

$$\Delta = 2 \times -0.22 \times 0.59/0.83 = -0.31.$$

Hence, in line with our first hypothesis, students with the same mathematics achievement had higher mathematics self-concepts if

Table 1
Predicting Mathematics Self-Concept (Study 1): Results From Multilevel Modeling

Fixed effects	Model 1			Model 2			Model 3		
	B	*p*	*SE*	*B*	*p*	*SE*	*B*	*p*	*SE*
Intercept	0.00	.955	.02	0.00	.946	.02	0.00	.880	.02
Class level									
Perceived class standing				0.21	<.001	.02	0.14	<.001	.02
Class-average achievement	−0.22	<.001	.03				−0.28	<.001	.03
Individual level									
Perceived class standing				0.19	<.001	.02	0.17	<.001	.02
Mathematics achievement	0.60	<.001	.02				0.60	<.001	.02
Variance explained		.29			.04			.31	
Deviance		11,962.30			13,428.24			11,810.39	
Estimated parameters		6			6			8	

their class showed low average achievement than if their class showed high average achievement. The overall proportion of variance explained in self-concept was 29%. The model fit of Model 1 was statistically significantly better than that of an empty model (not reported in Table 1) with mathematics self-concept as the dependent variable, $\Delta\chi^2$ (2, N = 4,810) = 1,615.87, $p < .001$.

Our second hypothesis stated that there would be reliable between-class differences in students' perceptions of the standing of their mathematics classes. How strongly did the perceptions of students in the same class covary? Consensus (or, more correctly, reliability of student responses) is easily calculated by means of the intraclass correlation coefficients (ICC), ICC_1 and ICC_2 (see Bliese, 2000; Lüdtke, Trautwein, Kunter, & Baumert, 2006; Snijders & Bosker, 1999). The ICC_1 indicates the proportion of the total variance that is located between school classes; given the same total variance, the higher the ICC_1, the more similar the perceptions of the students in the same classes regarding the standing of their class. In our study, the ICC_1 amounted to .47, indicating that there were considerable between-class differences and within-class agreement in how class standing was perceived. The ICC_2 can be used to evaluate the reliability of the aggregated student ratings at the class level. It is a function of the ICC_1 and the number of students per classes. As a rule of thumb, ICC_2 values above .70 are seen as indicating sufficient reliability (Lüdtke et al., 2006). In the present study, with an average of 8.79 students per class, the ICC_2 was .88, indicating that the aggregated perceived class standing score was a reliable indicator of perceived mathematics class standing across classrooms. The aggregated perception of class standing was positively correlated with class-average achievement (r = .31, $p < .001$); this moderately close association indicates that perceived class standing reflects more than the class-average student ability. Taken together, in line with our hypothesis, we found that classmates reported similar perceptions of their class's standing, although individual students within the same mathematics class differed to some degree in their evaluations.

Our third—and most central—hypothesis stated that these differential evaluations of the class standing would also be reflected in students' mathematics self-concept at both the class and the individual level. We specified two additional multilevel models to address this hypothesis (Models 2 and 3 in Table 1). In Model 2, we replaced the achievement predictors by perceived class standing at the individual level. Findings at the individual level showed that students who had a higher opinion of their class's standing reported higher mathematics self-concept than their classmates who had a lower opinion of their class's standing. Similarly, findings at the class level showed that mathematics self-concept was higher in classes in which the overall students rating of class standing was high. The effect size for this effect was $\Delta = 2 \times 0.21 \times 0.74/0.97 = 0.25$.

In Model 3, we reintroduced the achievement variables at both levels. Thus, this model examined whether the positive effects of high perceived class standing would still be observable when objective differences in standardized achievement were controlled. As shown in Table 1, all predictor variables were statistically significantly associated with mathematics self-concept. The beta coefficients for perceived class standing remained quite stable from Model 2 to Model 3. At the class level, the regression coefficient decreased to 0.14; however, with a reduced residual standard deviation in mathematics self-concept of $\sigma_e = 0.82$, the effect size for class-level perceived prestige remained stable at $\Delta = 0.25$. Taken together, in line with Hypothesis 3, and congruent with the idea that people use social comparison processes adaptively to enhance self-evaluations in natural learning environments, we found higher mathematics self-concept in classes in which the overall rating of class standing was high (class-level effect) as well as in individual students who rated the standing of their mathematics class more favorably than their classmates (student-level effect). The theoretical model upon which our predictions were based posits that the total effect of school-average ability on self-concept is the net effect of two opposite effects, namely a positive reflected-glory assimilation effect and a negative contrast effect. Consistent with these expectations, the introduction of the perceived standing of the class (Model 3) also resulted in a somewhat more negative effect of school-average ability.

Finally, we tested whether some individual students were specifically affected by class characteristics. To this end, a total of four interaction effects were specified and tested for statistical significance. In this analysis, class-average achievement was found to interact with individual achievement. We found a statistically significant regression coefficient of $B = 0.11$ ($p < .001$) for this interaction term, indicating that—although students at all levels of achievement were negatively affected by frame of reference effects—high-achieving students in the present sample were

somewhat less affected than low-achieving students. All other interaction effects failed to reach conventional levels of statistical significance (all $ps > .20$).

In sum, the findings of Study 1 confirmed our main hypothesis. Most importantly, students within a class had similar perceptions of the standing of their class, and the perceived standing of the class was positively related to students' self-concept, offsetting to some extent the negative effects of class-average ability. Furthermore, as indicated by the significant interaction between class-average achievement and individual achievement, high-achieving students were somewhat less affected by the negative frame of reference effect than were low-achieving students.

Study 2

Method

Sample

Study 2 was essentially a replication of Study 1 with a different sample of students. The analyses in Study 2 are based on data from the Initial Achievement and Learning Development study (LAU; Lehmann, Vieluf, Nikolova, & Ivanov, 2006; Trautwein, Köller, Lehmann, & Lüdtke, 2007). In this study, all Grade 13 students in the German state of Hamburg took a mandatory achievement test in 2005. As an optional part of the study, they were invited to answer a questionnaire including a self-concept inventory and questions relating to the perceived standing of their mathematics classes. Only students who were enrolled in comparable mathematics classes (i.e., general rather than advanced mathematics classes) and who provided valid data were included in the present analyses, giving a sample of 1,502 students (58% female, mean age: $M = 19.8$ years, $SD = 1.10$) from 192 general mathematics classes in 72 schools.

Instruments

We used the same instruments as in Study 1. Internal consistency of the scale scores was .87 for mathematics achievement, .89 for mathematics self-concept, and .87 for perceived mathematics class standing.

Results and Discussion

We conducted the same set of analyses as in Study 1. Results from multilevel modeling are reported in Table 2. In Model 1, the negative frame of reference effect was replicated. Controlling for individual mathematics achievement, we again found that class-average mathematics achievement negatively predicted mathematics self-concept. With a standard deviation in class-average achievement of $SD = 0.55$ and a residual standard deviation in mathematics self-concept of $\sigma_e = 0.81$, the effect size was a sizeable $\Delta = -0.63$. The model fit of Model 1 was statistically significantly better, $\Delta\chi^2$ (2, $N = 1{,}502$) = 609.10, $p < .001$, than that of an empty model ($\chi^2 = 4{,}261.56$; not reported in Table 2) with mathematics self-concept as the dependent variable.

We next examined the consistency of classmates' perceptions of class standing by calculating the intraclass correlations coefficients. The ICC_1 was .39 and the ICC_2 was .83, indicating that there was considerable consistency in perceived class standing among classmates and justifying the use of this variable as a class-level construct. The positive association between the class-average perception of class standing and class-average achievement was nonsignificant ($r = .14$, $p = .06$).

In Model 2, we used individual and class-average perceptions of class standing to predict mathematics self-concept. Findings at the class level showed that students in classes with a high average perception of the standing of their class reported comparatively high self-concepts. With a standard deviation in class-average perceived standing of $SD = 0.71$ and a residual standard deviation in mathematics self-concept of $\sigma_e = 0.95$, the effect size was a sizeable $\Delta = 0.25$. Similarly, at the level of the individual student, we found comparatively high mathematics self-concept in students whose evaluations of class standing were higher than those of their classmates.

We simultaneously introduced all predictor variables in Model 3. Although the beta coefficients of the predictor variables decreased slightly, the pattern of results remained virtually the same. Due to the reduced residual standard deviation of mathematics self-concept in Model 3 ($\sigma_e = 0.79$), the effect sizes for class-average achievement ($\Delta = -0.64$) and aggregated perceived class standing ($\Delta = 0.27$) were almost unchanged.

Table 2
Predicting Mathematics Self-Concept (Study 2): Results From Multilevel Modeling

	Model 1			Model 2			Model 3		
Fixed effects	*B*	*p*	*SE*	*B*	*p*	*SE*	*B*	*p*	*SE*
Intercept	0.04	.091	.02	0.00	.955	.02	0.04	.061	.02
Class level									
Perceived class standing				0.17	<.001	.03	0.15	<.001	.03
Class-average achievement	−0.46	<.001	.06				−0.45	<.001	.05
Individual level									
Perceived class standing				0.38	<.001	.04	0.27	<.001	.03
Mathematics achievement	0.65	<.001	.02				0.61	<.001	.02
Variance explained		.33			.09			.38	
Deviance		3,652.45			4,123.90			3,542.12	
Estimated parameters		6			6			8	

Finally, we again specified four interaction terms between class characteristics and individual student characteristics. In this random-slope model, we again found a statistically significant interaction between class-average achievement and individual achievement ($B = 0.08$, $p < .05$), suggesting that high-achieving students were less affected by the negative frame of reference effect than were low-achieving students.

Taken together, Study 2 closely replicated the findings of Study 1. Most important, the perceived standing of the mathematics class predicted mathematics self-concept over and above the predictive effects of mathematics achievement at both the individual and the class level. In fact, the effect sizes for the class-level indicators were somewhat stronger than in Study 1. This result might reflect the greater diversity of classes in Study 2 (conducted in a city with strong social disparities) than in Study 1 (conducted in a rather more homogeneous state). Additionally, we again found high-achieving students to be less affected by the negative frame of reference effect.

Study 3

Method

Sample

In Study 3, we focused the school level rather than the class level. The data for Study 3 were provided by a large, ongoing German study conducted by the Max Planck Institute for Human Development, Berlin (see Köller, Watermann, Trautwein, & Lüdtke, 2004). The analyses are based on data from 4,247 Grade 13 students (55.5% female, mean age: $M = 19.58$ years, $SD = .83$) in 149 randomly selected upper secondary schools in the state of Baden-Württemberg. Two trained research assistants administered materials in each school between February and May 2002. Students participated voluntarily, without any financial reward.[2]

Instruments

Mathematics achievement and mathematics self-concept. We again used the instruments administered in Study 1. Internal consistency of the scale scores was .89 for both mathematics achievement and mathematics self-concept.

Perceived school standing. In contrast to Studies 1 and 2, in Study 3 we tapped students' perceptions of the standing of their school rather than that of their mathematics class. Accordingly, the item stem read "Relative to other schools ___," and we replaced "mathematics class" with "school" in the four items. Other than this, the items were identical to those used in Studies 1 and 2. Internal consistency of the scale score (Cronbach's alpha) was .77.

Statistical Analyses

We again specified a set of multilevel models to predict mathematics self-concept, this time using the school as the Level 2 unit. Accordingly, we computed school-average achievement as well as an aggregate of students' perceptions of school standing as school-level variables to be included in the two-level multilevel models.

Results and Discussion

The results for Study 3 are reported in Table 3. In Model 1, the negative frame of reference effect was replicated. Controlling for individual student achievement, we found a statistically negative predictive effect of school-average achievement, indicating that students with the same individual achievement levels reported lower mathematics self-concept if placed in higher achieving schools. Hence, our first hypothesis was supported. With a standard deviation in school-average achievement of $SD = 0.45$ and a residual standard deviation in mathematics self-concept of $\sigma_e = 0.86$, the effect size amounted to $\Delta = -0.31$. The model fit of Model 1 was statistically significantly better, $\Delta\chi^2$ (2, $N = 4{,}247$) = 2,791.86, $p < .001$, than that of an empty model ($\chi^2 = 13{,}619.50$; not reported in Table 3) with mathematics self-concept as the dependent variable.

We next calculated the reliability of the aggregated school standing measure. With an ICC_1 of .28, the intraclass correlation of perceived school standing was somewhat lower than the ICC of the mathematics class standing measures used in Studies 1 and 2, indicating somewhat more variation in classmates' evaluations of school standing than in their evaluations of class standing. However, in support of our second hypothesis, because an average of 28.5 students per school participated in the study, the aggregated school-level standing indicator was highly reliable, with an ICC_2 of .92. The correlation between perceived school standing aggregated at the school level and aggregated mathematics achievement was not statistically significant ($r = .13$, $p = .11$).

In Model 2, we related perceived school standing to mathematics self-concept at the school and student levels. Findings at the student level showed that perceived school standing was positively associated with mathematics self-concept. Students who evaluated their school's standing more favorably than their fellow students had a higher mathematics self-concept. At the school level, the association between perceived school standing and mathematics self-concept was not statistically significant. In other words, although there was agreement among students on the standing of their school, these shared perceptions were not reflected in their mathematics self-concepts.

In Model 3, mathematics achievement and perceived school standing were simultaneously entered in the regression equation. The association between mathematics achievement and mathematics self-concept was positive at the student level and negative at the school level. Furthermore, perceived school standing was positively associated with mathematics self-concept at the student level, but not the school level. Hence, somewhat unexpectedly, we found only partial support for our third research hypothesis in this study.

Finally, we specified the four interaction terms between school-average achievement and school-average perception of school standing, on the on hand, and individual achievement and individual perception of school standing, on the other. In this random-slope model, we again found a statistically significant interaction between school-average achievement and individual achievement ($B = 0.13$, $p < .05$), suggesting that high-achieving students were less affected by the negative frame of reference effect than were

[2] About half of the students in the present sample were also examined in a study by Marsh et al. (2007), who investigated the long-term stability of frame of reference effects. However, Marsh et al. did not include the measure of school standing considered here.

Table 3
Predicting Mathematics Self-Concept (Study 3): Results From Multilevel Modeling

Fixed effects	Model 1			Model 2			Model 3		
	B	*p*	*SE*	*B*	*p*	*SE*	*B*	*p*	*SE*
Intercept	0.00	.861	.02	0.00	.899	.02	0.00	.855	.02
School level									
Perceived school standing				0.08	.077	.04	0.04	.233	.04
School-average achievement	−0.23	<.001	.04				−0.24	<.001	.04
Individual level									
Perceived school standing				0.06	.001	.02	0.07	<.001	.02
Mathematics achievement	0.53	<.001	.02				0.53	<.001	.02
Variance explained		.25			.00			.25	
Deviance		10,827.62			11,999.33			10,804.65	
Estimated parameters		5			5			7	

low-achieving students. The other three interaction effects did not reach conventional levels of statistical significance ($ps > .20$).

In sum, the findings of Study 3 replicated the negative frame of reference effect of school-average achievement and again suggested that high-achieving students are less affected by frame of reference effects than are low-achieving students. Moreover, students' individual perceptions of their school's standing were associated with their self-concept, whereas the aggregated indicator of the school's standing did not predict self-concept. This finding indicates that shared beliefs about between-school differences in a school's standing are not systematically integrated into students' self-concepts.

General Discussion

In three large empirical studies, we extended previous research on the negative frame of reference effects postulated by Marsh (1987) in educational settings. There were four main findings. First, given comparable individual achievement, placement in high-achieving learning groups is associated with comparatively low academic self-concepts. We found the same pattern of results for class-average and school-average achievement. Second, there was substantial agreement across students' perceptions of the relative standing of their class or school, yielding a reliable indicator of perceived class or school standing at the class or school level. Third, the students' academic self-concepts were not merely a reflection of their relative position within the class; they were also substantively associated with their perceptions of their class's overall standing. Across all three studies, student-level perceptions of class or school standing were positively associated with self-concept; furthermore, class-average student perceptions of class standing predicted self-concept. Hence, these findings are consistent with the idea that students actively integrate social comparison information into their self-concept (Suls & Wheeler, 2000). Our findings thus support key ideas from both educational and social psychology research on social comparison processes. Fourth, a statistically significant interaction effect in all three studies indicated that high-achieving students were less affected by the negative frame of reference effect than were low-achieving students.

Frame of Reference Effects: Generalizability and Moderator Effects

Students' placement in certain schools and classes can have major implications for their academic self-concepts. The complexity of real-life educational environments can hardly be modeled in laboratory experiments; rather, it is important to adopt a multilevel strategy at both the conceptual and the empirical level (Raudenbush & Bryk, 2002). In our studies, the adoption of a multilevel approach proved fruitful in several respects.

In line with previous educational psychology research on frame of reference effects (see Marsh & Craven, 2002), achievement proved to be differentially related to academic self-concept at the individual level and at the class level. Our studies thus add to the fairly consistent body of findings showing that placement in high-achieving learning environments is associated with relatively low academic self-concepts. The negative frame of reference effect was found whether the class or the school was used as the aggregate unit. Moreover, we extended prior research by calculating effect sizes for the frame of reference effects. The effect sizes observed were of meaningful magnitude, indicating that frame of reference effects indeed matter. Hence, unlike many social psychology studies highlighting powerful assimilation effects (see Mussweiler, 2003; Wheeler & Suls, 2005), we found that—when individual achievement is controlled—high-achieving classrooms may elicit contrast effects. The observed contrast effect is consistent with Diener and Fujita's (1997) observation that classrooms may act as "total environments" whose pervasive impact is not easily shaken off.

Despite the consistency of the frame of reference effect across the three studies and its meaningful effect size, we also found some indication that the total-environment-like nature of the frame of reference effect does not apply to all students in exactly the same way. More specifically, we found that the academic self-concepts of high-achieving students in all three samples were somewhat less influenced by the negative frame of reference effect than were the self-concepts of low-achieving students. From a theoretical point of view, this finding is quite reasonable. Put simply, high-achieving students may have less reason to be afraid of high-achieving classmates because they still fare quite well in a socially competitive arena. Nevertheless, given the inconclusive results of

earlier studies probing for this interaction effect, the consistent empirical support for this positive cross-level interaction across all three studies was somewhat unexpected (see Marsh, Seaton, et al., 2008). We can only speculate about possible reasons for the consistent findings found in our study. The most important factor might be the characteristics of the sample. Compared with other BFLPE studies, the students in the present research, who were in their last year of high school education, were relatively old. Moreover, because only students in the academic track were included (in Germany, all other students leave school after Grade 10), the samples were more selective. It seems possible that the interaction effect that we found is specific to such selective samples. Moreover, it seems important to note that we did not control for any mediator variables when testing the interaction effects. Including mediator variables such as school grades or student perceptions of teaching characteristics might negatively influence the strength of the interaction effect (e.g., Lüdtke et al., 2005; Marsh et al., 2007). Finally, the overall size of the interaction effect was relatively small; accordingly, the size of the frame of reference effects was affected, but not their overall direction. To conclude, more research is warranted on the conditions under which high individual achievement buffers the impact of class- or school-average achievement. This research requires not only substantive theoretical hypotheses, but also large sample sizes and the careful specification and testing of statistical models.

Integrating Class or School Standing in Self-Concept Research

A large number of studies have yielded evidence for negative frame of reference effects (see Marsh, Seaton, et al., 2008). It would be wrong to deny the occurrence of any adaptive social comparison processes in classroom settings, however. Studies teasing apart frame of reference effects and reflected glory processes have typically relied on "objective" characteristics of the school, such as track status (e.g., Marsh et al., 2001). Although they have produced some evidence for reflected glory effects, this evidence is not conclusive. Reflected glory effects as measured by a questionnaire scale were first described and analyzed in detail in the study by Marsh et al. (2000), but this study did not separate individual and class- or school-level effects. Taking a multilevel perspective, our study extended this pioneering study and found support for other information processing processes that led to differential academic self-concept across classes or students.

Our study adds to the literature on reflected glory effects in two ways. First, we were able to distinguish between students' individual perceptions of class or school standing and their shared perceptions of the standing of their learning environment. The perception of an individual student is an important source of information in social comparison studies, but individual ratings are prone to person-specific biases. For this reason, we also used aggregated individual perceptions as a measure of shared beliefs about the standing of the class or school. The analyses indicated that students' perceptions of the standing of their class or school are not highly idiosyncratic; rather, they reflect shared beliefs that distinguish between educational contexts. Social comparison processes evidently take place at the individual level and the group level. Students think about the characteristics of groups of students and use this information to evaluate their own academic qualities. Moreover, the students in a class share remarkably similar beliefs about their own class and other classes. Taken together, by adopting a multilevel perspective, our research went beyond the classical research design (Suls & Wheeler, 2000; Wood & Wilson, 2003) in which one perceiver rates or selects one specific target. As our analyses indicated, the ratings of students in the same class or school provide a highly reliable indicator of the perceived standing of the class when aggregated to the class or school level.

Although the scale we used to measure perceived class or school standing was rather short, it evidenced high psychometric quality. Factor and reliability analyses confirmed it to be unidimensional and internally consistent. In addition, the high reliability of the construct at the class level indicates that the scale taps an aspect of classroom or school reality that is relevant to students and that is relatively easy to describe in rating scales.

Still, some questions remain. What exactly does the perceived class or school standing construct measure? What is the likely genesis of students' shared beliefs about the standing of a class or school? Our four items tap the perceived amount learnt in the class or school, the overall achievement level of students in the class or school, and more general evaluations of how the class is perceived by others. We prefer the term *perceived standing* to some alternative candidates,[3] such as *class or school prestige* or *class or school status* (see Marsh et al., 2000). The term *prestigious* is typically used to describe selective schools or school tracks that admit students on the basis of their high achievement or other merits. In tracked school systems, for instance, the highest tracks have the most prestige (Trautwein, Lüdtke, Marsh, et al., 2006). In our study, however, all students attended the same (precollege) track; furthermore, their placement in a specific class or school was not determined by prior achievement. Hence, prestige or status may not be the most appropriate terms in the present context. Another alternative would be *perceived achievement*. However, perceived achievement emphasizes achievement at a certain time point, whereas one of our items emphasizes the learning trajectory. Moreover, despite the positive correlations between class-average achievement and class-average perceptions of the standing of the class, which indicate that students' shared perceptions reflect true differences in class achievement, the size of the correlation was only small to moderate. Whereas perceived achievement might be a too narrow term, another alternative—*perceived class or school quality*—might be too broad, because quality encompasses more than achievement growth and relative standing. Taken together, we believe the term *perceived standing of the class or school* adequately reflects the nature of the construct. However, more research is warranted to explore the predictors of high perceived class or school standing. Moreover, given that school systems across the world differ markedly, it is quite possible that studies conducted in other countries would use constructs that differ from the one used in the present study.

The second important contribution of our study to the analysis of reflected glory effects is the finding that perceptions of the standing of the class or school matter for academic self-concept at both the individual and the aggregate level. Hence, we found support for information processing processes leading to differen-

[3] We thank one anonymous reviewer for important feedback on the content and labeling of the perceived class or school standing construct.

tial academic self-concept across classes or students. Our studies help to distinguish between processes at the individual, class, and school levels. At the individual level, perceived class or school standing was statistically significantly related to academic self-concept in all three studies. At the class level (Studies 1 and 2), we also found a positive association between perceived class standing and achievement. At the school level (Study 3), however, this association was not statistically significant.

Previous studies have found students' academic self-concepts to be only slightly associated or not at all associated with the status of a school (Marsh et al., 2001; Trautwein, Lüdtke, Marsh, et al., 2006; Schwarzer et al., 1982), implying that students do not systematically integrate social comparison information about the standing of their school into their self-concept. In line with earlier research, Study 3 found that students in a school generally perceived to have a high standing do not report higher academic self-concepts. In contrast, students in mathematics classes collectively perceived to have high standing do report higher academic self-concepts. Hence, different classes within the same school seem to constitute important frames of reference for students, whereas different schools seem to constitute less salient frames of reference in terms of reflected glory. This does not mean, however, that an individual student's perception of school standing is irrelevant. In fact, a student whose perception of the standing of his or her school or class was higher than that of his or her school- or classmates was likely to report a comparatively high mathematics self-concept (also see Marsh et al., 2000).

Taken together, students' self-concepts are likely the product of a complex net of social comparison information that is partly person specific and partly the result of shared perceptions. When students' self-concepts are studied in natural environments, it is imperative to distinguish between the various levels of analysis. The results of our studies indicate that students actively seek out information about their own standing and the standing of their class and integrate that information into their academic self-concepts. Although placement in a high-achieving learning environment generally has a negative impact on academic self-concept, a positive evaluation of that environment's standing tends to buffer the negative impact somewhat.

Limitations and Conclusion

Although our investigation was based on a well-established theoretical model and used three strong data sets, some potential limitations should be addressed in future studies. First, our studies were nonexperimental and used a single-measurement design, meaning that caution is warranted in making causal interpretations. In nonexperimental studies there is always the possibility that untested variables affected the pattern of results. Yet, in real-world situations that can have important implications for students' achievement, motivation, and educational careers, the possibilities for true random assignment of students to conditions are severely limited for ethical and legal reasons. Large studies with random samples of students and powerful analytical tools provide a strong alternative to the experimental approach.

Second, generalizability is an issue. It is unclear to what extent differences across school systems might affect the results. School-level differences (in contrast to class-level differences) might be stronger in some school systems than in others. For instance, in countries in which school achievement scores are published in the form of ranking charts, students may be more aware of between-school differences, which in turn may affect their academic self-concepts. If this were indeed the case, it would constitute a moderator effect at a rather general level. Hence, we would like to see future studies replicate our analyses in diverse samples.

Third, our study focused on self-concept as the outcome variable. Given that a host of studies have demonstrated the importance of domain-specific self-concepts for short-term and long-term academic motivation (Marsh et al., 2005, 2007), academic effort (Trautwein & Lüdtke, 2007; Trautwein, Lüdtke, Schnyder, & Niggli, 2006), and academic choices (Watt & Eccles, 2008), this choice is justified. Nevertheless, additional outcome variables should be included in future studies.

The present study contributed to a research field of high practical importance and theoretical interest by taking a multilevel approach and including indicators of perceived class and school standing. The results provide further support for the powerful impact of frame of reference effects in educational environments. At the same time, our findings indicate that self-concept is not just the result of individual achievement moderated by a frame of reference effect, as readers of studies in educational psychology might be tempted to reason, but the result of a complex social comparison process involving several sources of information.

References

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: Freeman.

Baumert, J., Bos, W., & Lehmann, R. (Eds.). (2000). *TIMSS/III: Dritte Internationale Mathematik- und Naturwissenschaftsstudie: Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* [TIMSS/III: Third international mathematics and science study: Students' knowledge of mathematics and science at the end of secondary education]. Opladen, Germany: Leske & Budrich.

Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2008). *The effect of schooling on psychometric intelligence: Does school quality make a difference?* Manuscript submitted for publication.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco: Jossey-Bass.

Cialdini, R. B., & Richardson, K. D. (1980). Two indirect tactics of image management: Basking and blasting. *Journal of Personality and Social Psychology, 39,* 406–415.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coleman, J. M., & Fults, B. A. (1985). Special class placement, level of intelligence, and the self-concept of gifted children: A social comparison perspective. *Remedial and Special Education, 6,* 7–11.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulations of questions, design and analysis.* Stanford, CA: Stanford Evaluation Consortium.

Dai, D. Y., & Rinn, A. N. (2008). The big-fish-little-pond effect: What do we know and where do we go from here? *Educational Psychology Review, 20,* 283–317.

Damisch, L., Mussweiler, T., & Plessner, H. (2006). Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied, 12,* 166–178.

Diener, E., & Fujita, F. (1997). Social comparison and subjective well-being. In B. P. Buunk & F. X. Gibbons (Eds.), *Health, coping, and*

well-being: Perspectives from social comparison theory (pp. 329–358). Mahwah, NJ: Erlbaum.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12,* 121–138.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117–140.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* Mahwah, NJ: Erlbaum.

James, W. (1963). *The principles of psychology.* New York: Holt, Rinehart & Winston. (Original work published 1890)

Köller, O., Watermann, R., Trautwein, U., & Lüdtke, O. (2004). *Wege zur Hochschulreife in Baden-Württemberg. TOSCA – Eine Untersuchung an allgemein bildenden und beruflichen Gymnasien* [Educational pathways to college in Baden-Württemberg: TOSCA—A study at upper secondary level of traditional and vocational *Gymnasium* schools]. Opladen, Germany: Leske & Budrich.

Karabenick, S. A. (2004). Perceived achievement goal structure and college student help seeking. *Journal of Educational Psychology, 96,* 569–581.

Klusmann, U., Kunter, M., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Teachers' well-being and the quality of instruction: The important role of self-regulatory patterns. *Journal of Educational Psychology, 100,* 702–715.

Lüdtke, O., Köller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish-little-pond effect. *Contemporary Educational Psychology, 30,* 263–285.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing features of the learning environment: How to use student ratings in multilevel modeling. *Contemporary Educational Psychology, 34,* 120–131.

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research, 9,* 215–230.

Lehmann, R. H., Vieluf, U., Nikolova, R., & Ivanov, S. (2006). *LAU 13. Aspekte der Lernausgangslage und Lernentwicklung – Klassenstufe* 13 [LAU 13: Aspects of initial achievement and learning development]. Hamburg, Germany: Behörde für Bildung und Sport, Amt für Bildung.

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality. *Child Development Perspectives, 2,* 99–106.

Marsh, H. W. (1984). Self-concept: The application of a frame of reference model to explain paradoxical results. *Australian Journal of Education, 28,* 165–181.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79,* 280–295.

Marsh, H. W. (1991). The failure of high-ability high schools to deliver academic benefits: The importance of academic self-concept and educational aspirations. *American Educational Research Journal, 28,* 445–480.

Marsh, H. W., Chessor, D., Craven, R. G., & Roche, L. (1995). The effects of gifted and talented programs on academic self-concept: The big fish strikes again. *American Educational Research Journal, 32,* 285–319.

Marsh, H. W., & Craven, R. (2002). The pivotal role of frames of reference in academic self-concept formation: The big-fish-little-pond effect. In F. Pajares & T. Urdan (Eds.), *Adolescence and education* (Vol. 2, pp. 83–123). Greenwich, CT: Information Age.

Marsh, H. W., & Hau, K. T. (2003). Big-fish-little-pond effect on academic self-concept: A crosscultural (26 country) test of the negative effects of academically selective schools. *American Psychologist, 58,* 364–376.

Marsh, H. W., Köller, O., & Baumert, J. (2001). Reunification of East and West German school systems: Longitudinal multilevel modeling study of the big-fish-little-pond effect on academic self-concept. *American Educational Research Journal, 38,* 321–350.

Marsh, H. W., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology, 78,* 337–349.

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish-little-pond effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review, 20,* 319–350.

Marsh, H. W., & Shavelson, R. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist, 20,* 107–125.

Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). Big-fish-little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal, 44,* 631–669.

Marsh, H. W., Trautwein, U., Lüdtke, O., & Köller, O. (2008). Social comparison and big-fish-little-pond effects on self-concept and efficacy perceptions: Role of generalized and specific others. *Journal of Educational Psychology, 100,* 510–524.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76,* 397–416.

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review, 110,* 472–489.

Oakes, J. (1985). *Keeping track: How schools structure inequality.* New Haven, CT: Yale University Press.

Organization for Economic Cooperation and Development (2001). *Knowledge and skills for life: First results from the OECD Programme for International Student Assessment (PISA) 2000.* Paris: Author.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y., & Congdon, R. T. (2004). *HLM 6: Hierarchical linear modeling.* Chicago: Scientific Software International.

Reuman, D. A. (1989). How social comparison mediates the relation between ability-grouping practices and students' achievement expectancies in mathematics. *Journal of Educational Psychology, 81,* 178–189.

Rheinberg, F., & Enstrup, B. (1977). Self-concept of intelligence with pupils in special and normal schools: A group effect. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 9,* 171–180.

Roberts, B. W., Caspi, A., & Moffitt, T. (2003). Work experiences and personality development in young adulthood. *Journal of Personality and Social Psychology, 84,* 582–593.

Rost, J. (1996). *Testtheorie und Testkonstruktion* [Test theory and test construction]. Bern, Switzerland: Huber.

Schwanzer, A. D., Trautwein, U., Lüdtke, O., & Sydow, H. (2005). Entwicklung eines Instruments zur Erfassung des Selbstkonzepts junger Erwachsener [Development of a questionnaire on young adults' self-concept]. *Diagnostica, 51,* 183–194.

Schwarzer, R., Lange, B., & Jerusalem, M. (1982). Selbstkonzeptentwicklung nach einem Bezugsgruppenwechsel [Self-concept development after a reference-group change]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 14,* 125–140.

Seaton, M., Marsh, H. W., Dumas, F., Huguet, P., Monteil, J. M., Regner, I., et al. (2008). In search of the big fish: Investigating the coexistence of the big-fish-little-pond effect with the positive effects of upward comparison. *British Journal of Social Psychology, 47,* 73–103.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Validation of construct interpretations. *Review of Educational Research, 46,* 407–441.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

Suls, J., & Wheeler, L. (2000). *Handbook of social comparison: Theory and research.* New York: Kluwer Academic/Plenum Press.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103,* 193–210.

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 181–227). San Diego, CA: Academic Press.

Trautwein, U., & Baeriswyl, F. (2007). Wenn leistungsstarke Klassenkameraden ein Nachteil sind: Referenzgruppeneffekte bei Übergangsentscheidungen [When high-achieving classmates put students at a disadvantage: Reference group effects at the transition to secondary schooling]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 21,* 119–133.

Trautwein, U., Gerlach, E., & Lüdtke, O. (2008). Athletic classmates, physical self-concept, and free-time physical activity: A longitudinal study of frame of reference effects. *Journal of Educational Psychology, 100,* 988–1001.

Trautwein, U., Köller, O., Lüdtke, O., & Baumert, J. (2005). Student tracking and the powerful effects of opt-in courses on self-concept: Reflected-glory effects do exist after all. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *New frontiers for self research: Vol. 2. International advances in self research* (pp. 307–327). Greenwich, CT: Information Age.

Trautwein, U., Köller, O., Lehmann, R., & Lüdtke, O. (Eds.). (2007). *Schulleistungen von Abiturienten: Regionale, schulformbezogene und soziale Disparitäten* [School achievement of students at *Gymnasium* schools: Regional, track-specific, and social disparities]. Münster, Germany: Waxmann.

Trautwein, U., & Lüdtke, O. (2007). Students' self-reported effort and time on homework in six school subjects: Between-student differences and within-student variation. *Journal of Educational Psychology, 99,* 432–444.

Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Self-esteem, academic self-concept, and achievement: How the learning environment moderates the dynamics of self-concept. *Journal of Personality and Social Psychology, 90,* 334–349.

Trautwein, U., Lüdtke, O., Kastens, C., & Köller, O. (2006). Effort on homework in Grades 5 through 9: Development, motivational antecedents, and the association with effort on classwork. *Child Development, 77,* 1094–1111.

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth grade mathematics. *Journal of Educational Psychology, 98,* 788–806.

Trautwein, U., Lüdtke, O., Schnyder, I., & Niggli, A. (2006). Predicting homework effort: Support for a domain-specific, multilevel homework model. *Journal of Educational Psychology, 98,* 438–456.

Tymms, P. (2001). A test of the big fish in a little pond hypothesis: An investigation into the feelings of seven-year-old pupils in school. *School Effectiveness & School Improvement, 12,* 161–181.

Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educational research* (pp. 55–66). London: National Foundation for Educational Research.

Watt, H. M. G., & Eccles, J. S. (Eds.). (2008). *Explaining gendered occupational outcomes: Examining individual and social explanations through school and beyond.* Washington, DC: American Psychological Association.

Wheeler, L., & Suls, J. (2005). Social comparison and self-evaluation of competence. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 566–578). New York: Guilford Press.

Wheeler, L., & Suls, J. (2007). Assimilation in social comparison: Can we agree on what it is? *Revue Internationale de Psychologie Sociale, 20,* 31–51.

Wood, J. V., & Wilson, A. E. (2003). How important is social comparison? In M. L. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 344–366). New York: Guilford Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest: Generalised item response modelling software manual.* Melbourne, Australia: Australian Council for Educational Research.

Zeidner, M., & Schleyer, E. J. (1999). The big-fish-little-pond effect for academic self-concept, test anxiety, and school36 grades in gifted children. *Contemporary Educational Psychology, 24,* 305–329.

Received April 7, 2008
Revision received April 20, 2009
Accepted April 24, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 867–878

0022-0663/09/$12.00 DOI: 10.1037/a0016644

Intergenerational Family Predictors of the Black–White Achievement Gap

Jelani Mandara, Fatima Varner, Nereira Greene, and Scott Richman
Northwestern University

The authors examined intergenerational family predictors of the Black–White achievement gap among 4,406 adolescents from the National Longitudinal Survey of Youth. An intergenerational model of the process by which family factors contribute to the achievement gap was also tested. The results showed that the ethnic gaps in socioeconomic status (SES) and achievement had significantly reduced over the past few generations. Moreover, measures of grandparent SES, mothers' achievement, parent SES, and a comprehensive set of reliable parenting practices explained all of the ethnic differences in achievement scores. Parenting practices such as creating a school-oriented home environment, allowing adolescents to make decisions, and not burdening them with too many chores had particularly important effects on the achievement gap. The authors conclude that adjusting for these differences would eliminate the ethnic achievement gap.

Keywords: achievement gap, parenting, intergenerational

The ethnic variation in achievement has been a contentious topic for several decades (Jencks & Phillips, 1998). Studies of school-based factors show that the ethnic gaps in achievement are associated with ethnic differences in academic tracking (Oakes, 2005), teacher expectations (Ferguson, 2003), and the overall quality and quantity of educational experiences (Campbell, Pungello, Miller-Johnson, Burchinal, & Ramey, 2001; Fryer & Levitt, 2004). A study of data from the National Assessment of Educational Progress found that almost 40% of the achievement gap in math can be explained by the amount and type of math courses taken, rigor of curriculum, frequency of calculator use, and the degree to which students believe that math is fact learning (Byrnes, 2003). However, there has been less focus in educational research on the importance of parents' role in the achievement gap and their children's achievement generally. Consequently, the research on the family factors that may account for the achievement gap is less comprehensive.

Many researchers argue that family factors such as socioeconomic status (SES; Brooks-Gunn, Klebanov, Smith, Duncan, & Lee, 2003; McLoyd, 1998), parents' education (Byrnes, 2003), and certain parenting practices (Bradley et al., 1989; Moore, 1986) account for some of the ethnic differences in achievement, but the remaining variance is due to innate genetic differences (Herrnstein & Murray, 1994). For instance, one study assessed economic resources and family environment factors and concluded that they explained up to two thirds of the achievement gap between Black and White 5- and 6-year-olds (Phillips, Brooks-Gunn, Duncan, Klebanov, & Crane, 1998). Phillips, Brooks-Gunn, et al. (1998) further concluded that 26% of the family environment effect is genetic in nature.

However, there are significant methodological flaws in most previous studies of family factors that reduce the validity of the conclusions. For one, most previous studies tended to examine only a few parenting factors at a time—usually only parental warmth and/or cognitive stimulation in the home. Furthermore, many of the studies tended to use poorly measured parenting variables or single-item SES indicators. This can mask the true effects of the variables and lead to the conclusion that they do not account for much of the ethnic gap. Almost all of the large-scale achievement gap studies that assessed family factors have examined only young children. Given that ethnic gaps in achievement have been shown to increase as youths get older (Phillips, Crouse, & Ralph, 1998), failure to understand how family factors affect the gap during adolescence is a major shortcoming. Another limitation of previous studies is that they rarely examined the long-term impact of past generations' economic and social resources on the recent achievement gap. Only the Phillips et al. (1998) study described above examined grandparent factors. Probably the major limitation of prior studies is that little is known about the family processes that link differences in ethnic background to ethnic differences in achievement.

The purpose of the current study was to address these limitations and determine how much of the ethnic gap in adolescent achievement test scores can be explained by using a comprehensive set of relatively well-measured family factors. A further purpose was to test an intergenerational model of the process by which generations of SES and parenting differences may lead to ethnic differences in adolescent achievement. Specifically, data from the Na-

Jelani Mandara, Fatima Varner, Nereira Greene, and Scott Richman, Program in Human Development and Social Policy, School of Education and Social Policy, Northwestern University.

The National Longitudinal Survey of Youth and the Children of the National Longitudinal Survey of Youth surveys are sponsored and directed by the U.S. Bureau of Labor Statistics and the National Institute for Child Health and Human Development. The surveys are managed by the Center for Human Resource Research at the Ohio State University, and interviews are conducted by the National Opinion Research Center at the University of Chicago. We thank Greg Duncan and Lindsay Chase-Lansdale for their review of earlier versions of this article.

Correspondence concerning this article should be addressed to Jelani Mandara, Program in Human Development and Social Policy, Northwestern University, 2120 Campus Drive, Evanston, Il 60208. E-mail: j-mandara@northwestern.edu

tional Longitudinal Survey of Youth (NLSY) and National Longitudinal Survey of Youth–Child Supplement (NLSY-C) were used to examine the effects of grandparent SES, parent achievement and SES, and parenting factors on the achievement gap between African American and European American adolescents.

Conceptual Model of Intergenerational Family Predictors of the Achievement Gap

The model tested in this study is illustrated in Figure 1. The model suggests that events in American history created rather large ethnic differences in family SES. These differences in SES are the initial link in the chain of events leading to current ethnic differences in achievement. This hypothesis stems from the consistent finding that components of SES such as parental education, occupation, income, and wealth are highly correlated with academic achievement (Brooks-Gunn, Klebanov, & Duncan, 1996; Crane, 1996; McLoyd, 1998; Orr, 2003). There are also large ethnic differences in SES that appear to coincide with the ethnic differences in test scores. For instance, recent census data show that Asian Americans' average total household income was $44,080, European Americans' was $40,212, Hispanic Americans' was $30,291, and African Americans' was $26,168 (DeNavas-Walt, Cleveland, & Webster, 2003). In addition, African American children are not only more likely to live in poverty but also more likely to live in persistent poverty (Duncan, Brooks-Gunn, & Klebanov, 1994). Several researchers have therefore argued that differences in family SES account for much of the ethnic differences in test scores.

A few studies using the NLSY have found consistent support for this idea. Brooks-Gunn et al. (2003) used low-birth-weight 3- and 5-year-olds from the control group of a clinical trial study and 3- to 6-year-olds from the NLSY to assess the effects of family characteristics on the Black–White test score gap. Analyses of their tables show that income-to-needs ratio explained 16% to 50% of the ethnic differences in test scores beyond child characteristics, depending on the test, sample, and age of the children. Another study using the NLSY examined the impact of wealth on the Black–White test score gap (Orr, 2003). This is an important study because African American families have significantly less wealth than European American families, even when income, occupation, and parental education are controlled (Eller, 1994). Results showed that wealth explained an additional 15% of the test score gap, even after traditional measures of SES, maternal test scores, family structure, and other variables were controlled (Orr, 2003). Thus, various measures of family SES seem to account for significant portions of the ethnic test score gap.

The conceptual model in Figure 1 also suggests that the quality of parenting is the main mediator of the effects of SES on achievement. The higher the SES of parents, the more likely they are to use parenting practices associated with academic achievement. This assumption is consistent with the family stress model, which predicts that families' experience with economic pressure negatively impacts youths because it interferes with parents' mental health, which reduces the quality of parenting the youths receive (Conger et al., 2002). In support of this, several studies find that the effects of various SES factors on child and adolescent achievement are almost completely mediated by family functioning and parenting (Brody & Flor, 1997; Crane, 1996; Guo & Harris, 2000; Linver, Brooks-Gunn, & Kohen, 2002; Mistry, Vandewater, Huston, & McLoyd, 2002). For instance, Guo and Harris (2000) used

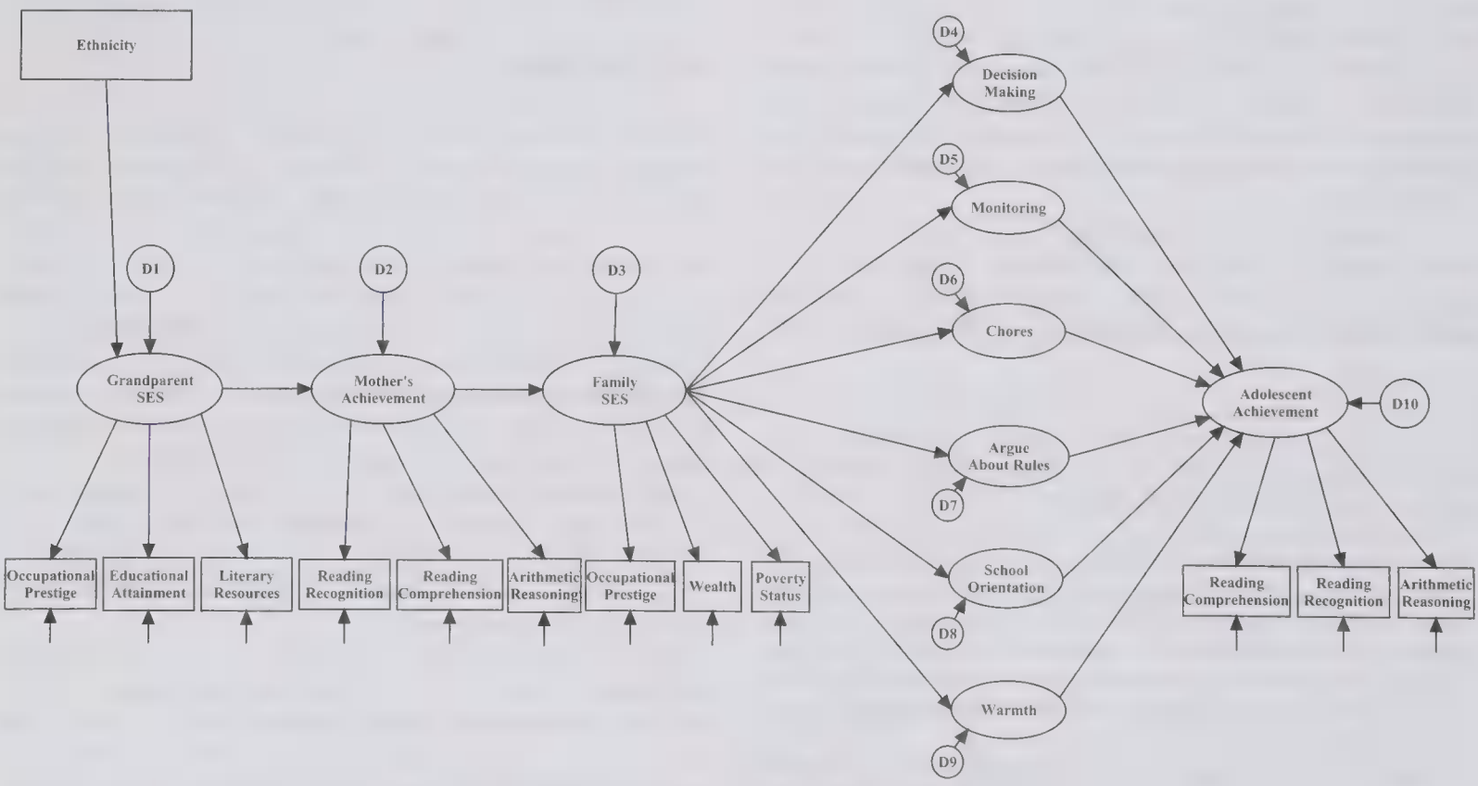


Figure 1. Conceptual model of intergenerational family factors on adolescent achievement. SES = socioeconomic status.

the NLSY and found that parental warmth, the physical setting of the home, and especially cognitive stimulation in the home completely mediated the effects of poverty on preschool children's cognitive development.

Parenting is also theorized to account for ethnic differences in achievement. This stems from the consistent findings that European American parents are rated by themselves and observers as enacting more practices consistent with achievement than African American parents. For instance, European American parents are higher in parental warmth and support (Bradley, Corwyn, Pipes McAdoo, & Garcia Coll, 2001; Moore, 1986), allow their children more autonomy to make decisions (Goldstein, Davis-Kean, & Eccles, 2005), and have a more cognitively stimulating home environment (Bradley et al., 2001; Brooks-Gunn et al., 1996), even after SES factors are controlled. In one of the most compelling studies, 23 African American children adopted by middle-class African American families were found to have an average IQ score of 103.6, whereas 23 others adopted by middle-class European American families had an average IQ of 117 (Moore, 1986). Mothers were also observed while helping their adopted children perform a difficult cognitive task. European American mothers were more likely than African American mothers to use positive reinforcement and other parenting strategies shown to be positively correlated with test scores (Moore, 1986). Each of these parenting practices is consistently related to academic achievement and cognitive development for European American and African American youth (Bradley et al., 2001; Brody & Flor, 1997; Brody, Stoneman, & McCoy, 1994; Linver et al., 2002).

Similarly, the research on parenting styles also clearly shows that European American youths are much more likely to have authoritative parents than African American youths (Steinberg, Mounts, Lamborn, & Dornbusch, 1991). Authoritative parents are warm and highly responsive to adolescents' emotional needs and allow autonomy of decision making, but they also monitor adolescents' friends and whereabouts, set high expectations, and have clear rules and behavioral boundaries (Baumrind, 1996). Studies consistently show that European American youths with authoritative parents outperform other European American adolescents at every level of education (Spera, 2005; Steinberg, Lamborn, Darling, Mounts, & Dornbusch, 1994; Strage & Brandt, 1999). Although early studies did not find a strong advantage for having authoritative parents among African American adolescents, as they did with European American adolescents, more recent studies do show that African American adolescents with authoritative parents have higher achievement and better mental health than those with authoritarian or other types of parents (Spera, 2005; Taylor, Hinton, & Wilson, 1995).

Only a few studies have directly tested the assumption that ethnic differences in parenting explain ethnic differences in achievement. One study using the low-birth-weight data set discussed above found that cognitive stimulation in the home and parental warmth explained 28% of the test score gap in preschool children, even after SES, maternal education, neighborhood conditions, and other factors were controlled (Brooks-Gunn et al., 1996). Brooks-Gunn et al. (2003) also found that these parenting variables explained 14% of the test score gap in the 3-year-olds of the NLSY and 16% of the gap for 5-year-olds, after SES and maternal achievement were controlled. Therefore, as these studies imply, the model in Figure 1 suggests that one reason European American adolescents have higher achievement scores than African American adolescents is that their parents have higher SES, which makes them more likely to use parenting strategies that foster academic achievement.

The conceptual model in Figure 1 also suggests that one consequence of differential family SES and parenting practices is that these resources tend to be transmitted from one generation to the next (Phillips et al., 1998; Saltaris et al., 2004). For instance, 12% of European Americans receive inheritances over $10,000, whereas only 1% of African Americans do (Gittleman & Wolff, 2002). European American parents are also more able to help their children financially with their education costs, home purchases, and payment of debt than African American parents are. This transmission of resources is one reason why European American families tend to have more wealth than African American families, even when parental education and incomes are equal (Gittleman & Wolff, 2002).

Grandparents may also have a direct impact on children because they often act as caregivers to their grandchildren. This may be particularly true for African Americans, considering the large amount of time they spend with their grandparents (Kamo, 2000). However, only one study to date has examined the effects of these factors on the test score gap in grandchildren (Phillips et al., 1998). In support of the conceptual model proposed here, they found that grandparents' resources explained about 25% of the test score gap before second generation factors were controlled. When the second generation factors were added to the model, grandparent effects were reduced. Thus, the model tested in this study suggests that European Americans' SES advantages from prior generations helps produce future generations' advantages in achievement through the middle generation's SES and parenting.

The Current Study

This study attempts to build on the strengths of prior studies and address their limitations by using a larger set of reliable SES and parenting factors from the 1979–2000 NLSY and NLSY-C data. The NLSY is a national panel study of over 12,000 women and men who began participating in 1979. At that time, participants were ages 14 to 21. The participants have been followed every year or every other year since 1979. In 1986 and every other year since, in-home assessments of the NLSY women's children were taken as part of the NLSY-C. Therefore, the NLSY and NLSY-C have data on grandparents (i.e., parents of the NLSY women), parents (i.e., the NLSY women), and adolescents (i.e., children of the NLSY women). These unique data allow for the assessment of SES factors in the mother's home when she was an adolescent. Unfortunately, grandparents' parenting practices and information about the adolescent's paternal grandparents were not assessed in the NLSY. The mother's achievement test scores when she was an adolescent were also assessed. As mothers grew up and had children, their families' SES and a comprehensive set of parenting practices were assessed, as well as the achievement test scores of their adolescents. To account for measurement error and test the conceptual model, structural equation modeling (SEM) methods were employed (Arbuckle, 2007; Bollen, 1989).

Given the literature reviewed above, we expected that the factors in the model would account for almost the entire ethnic achievement gap among the adolescents. Thus, the large differ-

ences in African American and European American adolescent achievement would be eliminated once the family factors were entered into the SEM. As depicted in Figure 1, we also hypothesized that the effects of grandparents' SES on adolescent achievement would be mediated by their daughters' achievement and SES. These factors would then be mediated by the parenting practices used in the adolescents' homes.

Method

Participants

The participants included 2,284 women from the original NLSY sample and 4,406 of their adolescents who were at least 10 years old in 2000 and who had data on the achievement variables. Age of adolescents was the criterion because most children at the age of 10 would be cognitively ready to answer questions about the parenting they received. To maintain the integrity of other design features of the NLSY-C, such as measuring grandparent factors before the children were born, those born before 1980 were not used in the current study. In 2000, the average age of the adolescents was 14.4 (SD = 2.3), and their mothers' average age was 39 (SD = 2.2). Fifty-one percent of the children were male. From 1986 to 2000, the average family income in 1999 dollars was $44,000 ($SD$ = $40,109). Ethnic group differences and other descriptive statistics are discussed in the Results section.

Procedure

In 1978, the National Opinion Research Center identified over 150,000 people from a list of housing units in selected areas of the United States. From this list, they identified a sample of over 11,000 noninstitutionalized individuals who were 14 to 21 by the end of 1978. The NLSY also oversampled African Americans so that better comparisons with European Americans could be made. Yearly 1-hr personal interviews of the respondents by trained personnel occurred from 1979 through 1994. Respondents have been interviewed every other year since 1994. Beginning in 1986, all the available children of the female respondents were assessed every other year with a variety of interview and survey methods. From ages 10–11 until 13–14, children were interviewed separately. Most parenting measures used in the current study derived from those interviews. The respondents were paid $10 for each interview from 1979 to 1994. They were paid $20 plus $5 per child for each assessment since 1994. The mothers were also paid $50 for the Armed Forces Qualification Test in 1980. Chase-Lansdale, Mott, Brooks-Gunn, and Phillips (1991) discussed the methods in more detail.

Instruments

Various measures of grandparents' SES, mother's achievement, the immediate family's SES, six parenting practices, and adolescent achievement were assessed. Grandparent factors were assessed when the mother was in the first year of the NLSY study in 1979–1980. Mother's achievement was assessed in 1979–1980 before the adolescents were born. Family SES was derived from the mother's self-reports as she became an adult. From the time a child turned 10–11 until the age of 13–14, he or she was interviewed separately once and sometimes twice between those years. The parenting scores were averaged for those with two assessments. Other parenting variables were derived from mothers' self-reports and observations of trained staff. Factor and reliability analyses were conducted to create each variable. The data file was also split by ethnicity, and the final results were repeated to verify equivalence across ethnic groups. Each of the variables had roughly equal factor loadings and reliabilities for African Americans and European Americans. Due to space limitations, the details of the factor analyses are not presented here. The coefficient alphas and other measurement details are presented below.

Grandparents' educational level. During the first year of the study, the mothers were asked the highest levels of education completed by their mother and their father (if known) by 1979. A composite of grandparents' education was then created by averaging grandmother's and grandfather's highest levels of education (α = .70).

Grandparents' occupational prestige. Interviewers assessed the occupations of the grandmothers and grandfathers in the study by asking their daughters four questions. The first questions asked each daughter (i.e., mother in the current study) to name the occupations of her mother in 1979 and her father in 1979 (if known). The daughters were also asked to state the occupations of the adult male in their home when they were 14 and the adult female in their home when they were 14 (if known). As a way to classify occupations and organize them according to a hierarchy of prestige, the NLSY used the Census Bureau's three-digit 1980 occupational classification system. The one to four separate three-digit codes were then averaged to form the grandparents' occupational prestige composite (α = .82). The scale was reversed so that higher scores represent more prestigious occupations.

Grandparents' literacy resources. Another indicator of grandparents' SES was the amount of reading material around the home when the mother was 14. This was estimated from three items. The first item asked whether someone in the family subscribed to newspapers. The second asked whether someone in the family subscribed to magazines. The last item asked whether someone in the family had a library card. A composite was created by averaging these three items (α = .56).

Mothers' achievement test scores. Mothers' achievement was estimated in 1980 with the Armed Forces Qualification Test raw scores. The test is composed of the arithmetic reasoning, reading recognition, and reading comprehension sections of the Armed Services Vocational Aptitude Battery. The mothers took the test in 1980. Cronbach's alpha for each subtest was around .90.

Parents' occupational prestige. The mothers were asked to state their occupation, if any, and their husband or partner's occupation, if any, during each assessment. The Census Bureau's three-digit 1980 occupational classification system was applied to the reported occupations. Mother's and father's three-digit occupational prestige scores were averaged at each assessment until the child was 12–14 years old (α = .85). The scale was reversed so that higher scores indicate more prestigious occupations.

Parents' poverty status. In the NLSY, family poverty status was determined using the poverty income guidelines from the U.S. Department of Health and Human Services. This index of poverty takes into consideration family income and family size and is adjusted for inflation annually. The factor used in the current study is the proportion of time the family did not live in poverty from the adolescent's birth until the age of 12–14.

Parents' wealth. During the yearly interviews, respondents were asked to estimate the value of their home if they were homeowners. The average value of the home the family lived in from the adolescent's birth to age 12–14 was used as the indicator of parents' wealth. Those who never owned their homes were given a value of $0.

Adolescent decision making. During the NLSY-C, adolescents were asked a series of questions with the stem, "Who usually makes the decisions about _____?" The questions were: (a) "buying your clothes," (b) "how to spend money," (c) "which friends to go out with," (d) "how late you can stay out," (e) "how much allowance you get," and (f) "how much TV you can watch." They were asked to circle responses indicating themselves, their mother, father, stepfather, friend and/or other person. For the purposes of this study, only those times in which the adolescents indicated themselves were counted as a "yes" response. If they did not indicate themselves, it was scored as a "no" response. These scores were averaged to create the variable (α = .63).

Parental monitoring. The degree to which parents monitored the whereabouts of their adolescents was assessed on the NLSY-C with three items. Using a 5-point scale, two items had the stem, "How much do you tell your parent(s) about _____?" The questions were, "where you are when you are not at home" and "who you are with when you are not at home." A third question asked, "About how often does each parent know who you are with when you're not at home?" These scores were averaged to create the variable (α = .71).

Household chores. Adolescents were also asked the stem, "In your home, are you regularly expected to help out with _____?" They answered *yes* or *no* to (a) "keeping the rest of the house clean," (b) "doing the dishes," and (c) "cooking." These scores were averaged to create the variable (α = .50).

Arguing about rules. Also from the NLSY-C, this variable was derived from a set of questions that used the stem, "How often do you argue with your parent(s) about the rules about _____?" The questions were (a) "watching television," (b) "keeping your parents informed about where you are," (c) "doing your homework," and (d) "dating and going to parties with boys and girls." A 3-point Likert-type scale (from 1 = *hardly ever* to 3 = *often*) was used. Factor and reliability analyses found that the questions about watching television and dating did not correlate well with the other two items or each other. The remaining two items about keeping parents informed and homework rules were averaged to form the variable. Higher scores mean hardly ever arguing about the rules (α = .70).

Maternal warmth. Starting in 1986 and continuing every other year, observers completed a short form of the Home Observation for Measurement of the Environment (Caldwell & Bradley, 1984) for each family. A total of four items that were concerned with the mother's warmth toward and support of the child (i.e., "Did the mother encourage the child to contribute to the conversation?") were rated by trained and experienced observers on a *yes* or *no* scale. The items were averaged to form the variable (α = .70).

School-oriented home. During the same home visits, mothers were asked three questions about the school orientation of their home environment. They were asked about the number of books in the home (*10 or more books*, *3 to 9 books*, *1 or 2 books*, *none*); they were asked, "Is there a musical instrument that the child can use here at home?" (*yes* or *no*); and, "Does the child get special lessons or belong to any organization that encourages activities such as sports, music, are, dance, drama, etc?" (*yes* or *no*). The items were averaged to form the variable (α = .53).

Adolescent achievement. Adolescent achievement was estimated with the reading recognition, reading comprehension, and mathematical reasoning subtests of the Peabody Individual Achievement Test (PIAT). The PIAT battery is a wide-range, brief assessment of academic achievement first developed in 1970. The latest revisions and restandardization of the PIAT were conducted after the initial 1986 assessments. The adolescents took the PIAT battery every other year, beginning in their first year in the study. They completed the PIAT in their homes with trained interviewers. For this study, only the assessments of each subtest between the ages of 10–11 and 13–14 were aggregated. The alpha for each of the subtests was around .90.

Analysis Plan

To determine if the effects of ethnicity could be explained by the family factors and to test the intergenerational model in Figure 1, we employed latent variable SEM with maximum likelihood estimation using Amos 16.0 (Amos Development Corporation, Spring House, PA). SEM with latent variables is advantageous because it allows for the measurement error to be modeled. Furthermore, SEM allows for a test of the statistical significance of the direct (i.e., nonmediated) and indirect (i.e., mediated) effects. A nonsignificant direct effect and a significant indirect effect of ethnicity would indicate that ethnic differences in adolescent achievement were completely explained by the other variables in the model. Bootstrap methods were used to estimate the standard errors and p values of the indirect effects (Arbuckle, 2007). Overall model fit was assessed with the comparative fit index (CFI), the adjusted goodness-of-fit index (AGFI) and the root-mean-square error of approximation (RMSEA). Established criteria suggest that a CFI and AGFI of .95 or better and an RMSEA less than .06 indicate an excellent fit of the model to the data (Hu & Bentler, 1999).

Because some mothers had multiple children in the study, there is the possibility that the standard errors could be underestimated due to a violation of the assumption of independence. In other studies with the NLSY this has not been a concern, because of the relatively large sample size. Nevertheless, to check for this possibility, the adolescents were divided into the first, second, and third or later birth order groups, and multigroup analyses were conducted on the final model. The results were virtually identical in each subsample. The achievement gap was slightly larger at first for the first-born adolescents, but after the final model results were the same. Therefore, only the full sample results are presented.

Imputation of Missing Data

During the collection of the NLSY data, some participants were not available during every assessment. However, many issues had to be considered before data could be imputed. The participants were different ages in each year, and assessments only occurred every other year. Also, the parenting variables were assessed no more than twice between the ages of about 10 to 14. To account for these issues, we converted each variable to the age of the participant at assessment. Variables were then collapsed into age groups: 0–2, 3–4, 5–6, 7–8, 9–11, and 12–14. Adolescents had to have at

least one assessment of achievement between the ages of 10 and 14 to be included in the study. Missing data were then analyzed with the Missing Value Analysis add-on module in SPSS 16.0. Little's test showed that the data were not missing completely at random, $\chi^2(67758, N = 4{,}406) = 75{,}877.91$, $p < .001$. We therefore imputed missing data with the expectation maximization (EM) algorithm. This method replaces missing values with iterative maximum likelihood estimations based on the available data (Arbuckle, 2007). Once data were imputed, the 12- to 14-year-old data for those who were younger than 12 during the final assessment were removed. The age groups were then aggregated as described in the Method section.

Results

Descriptive Statistics

The zero-order correlations of the study variables are presented in Table 1. As expected, all of the grandparent SES factors had moderate to large correlations with their daughters' achievement and family SES. Mothers' achievement and each of the family SES variables also had moderate to large correlations with the parenting variables in the expected directions. Adolescents whose mothers had higher achievement and SES experienced more autonomy in their decisions, more parental monitoring, warmer parent–adolescent relationships, and more school-oriented homes. As can be seen, adolescents' math and reading scores had significant correlations with all the other study variables. The school orientation of the home, adolescent freedom to make decisions, and the degree of arguing about the rules had particularly strong correlations for the parenting variables.

Table 2 presents the means, standard deviations, and independent sample t tests for each variable by ethnicity. The groups differed on all variables in typical directions. European American grandparents had significantly higher education, occupational status, and more literacy resources than African American grandparents. European American parents were also less likely to have experienced poverty and had more prestigious occupations and much more wealth. Although European American mothers and adolescents outscored African Americans on the achievement tests, as is typically found, the effect size of the differences dramatically decreased in this one generation. Whereas 23% to 29% of the variance in mothers' test scores could be explained by ethnicity, only 9% to 14% of their adolescents' test scores could be explained by ethnicity. In like manner, although European American parents were better off on all the SES variables than African American parents, the sizes of the ethnic differences in occupational prestige dropped from the grandparents' generation to the parents' generation at roughly the same rate as the size of the achievement test score gap reduced from the parents' to the adolescents' generation.

Measurement Model

In the measurement phase of the analyses, a confirmatory factor analysis was conducted that included only the latent variables with multiple indicators (i.e., grandparents' SES, mothers' achievement, family SES, and adolescents' achievement). To scale each latent variable, the paths to the indicators with the highest loading for each latent variable were set to 1 (Arbuckle, 2007). The latent variables were allowed to covary in the model. The constrained model yielded a good fit, $\chi^2(48, N = 4{,}406) = 1040.13$, $p < .001$, CFI = .97, AGFI = .96, RMSEA = .06. Also, the loadings and R^2 for each indicator were moderate to large (see Table 3).

The six parenting variables were used as single indicators of their latent variables in the analyses. To account for measurement error, the paths from the latent variables to the indicators were

Table 1
Zero-Order Correlations Between Study Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Grandparents																		
1. Occupational prestige	—																	
2. Education	.55	—																
3. Literary resources	.37	.45	—															
Parents																		
4. Arithmetic reasoning	.37	.41	.39	—														
5. Word recognition	.43	.47	.47	.73	—													
6. Reading comprehension	.37	.42	.44	.69	.82	—												
7. Occupational prestige	.28	.32	.25	.33	.33	.31	—											
8. Poverty	.38	.37	.38	.50	.60	.58	.21	—										
9. Wealth	.36	.36	.32	.46	.43	.40	.39	.46	—									
10. Decisions	.22	.23	.21	.30	.34	.34	.12	.30	.21	—								
11. Monitoring	.18	.23	.20	.22	.23	.23	.21	.24	.29	.09	—							
12. House chores	−.19	−.17	−.16	−.25	−.29	−.27	−.16	−.29	−.24	−.14	−.10	—						
13. Argue about rules	−.13	−.17	−.13	−.20	−.24	−.25	−.15	−.19	−.16	−.24	−.13	.12	—					
14. School orientation	.39	.40	.42	.49	.54	.50	.33	.56	.45	.30	.27	−.23	−.23	—				
15. Warmth	.15	.15	.16	.19	.21	.21	.14	.27	.17	.13	.10	−.06	−.10	.26	—			
Adolescents																		
16. Arithmetic reasoning	.32	.35	.31	.49	.49	.48	.31	.42	.37	.42	.25	−.28	−.33	.47	.19	—		
17. Word recognition	.28	.30	.29	.41	.48	.45	.28	.39	.32	.36	.20	−.23	−.40	.44	.16	.68	—	
18. Reading comprehension	.28	.32	.31	.46	.52	.48	.27	.43	.35	.38	.18	−.26	−.38	.48	.17	.71	.83	—

Note. $N = 4{,}406$. All rs are significant at $p < .001$.

Table 2
Means, Standard Deviations, and t Tests of Each Factor by Ethnicity

Factor	Black		White			
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	*t*	*d*
Grandparent SES						
Occupational prestige	316.80	200.75	490.34	219.78	27.00	0.82
Education	9.90	2.61	11.38	2.41	19.52	0.59
Literary resources	0.51	0.35	0.76	0.30	25.66	0.77
Mother's achievement						
Arithmetic reasoning	10.36	4.20	16.96	6.70	37.86	1.18
Word recognition	16.92	7.26	26.10	7.13	42.05	1.28
Reading comprehension	7.76	3.38	11.34	3.08	36.61	1.11
Family SES						
Occupational prestige	450.58	140.51	489.74	160.63	8.47	0.39
Poverty status	0.51	0.35	0.84	0.25	35.79	1.09
Wealth	9,344.38	29,416.75	62,940.85	77,992.11	28.63	0.91
Parenting practices						
Decision making	0.38	0.16	0.48	0.15	21.01	0.64
Parental monitoring	2.73	0.52	2.93	0.50	13.27	0.39
House chores	0.68	0.19	0.55	0.23	−20.00	0.62
Argue about rules	1.70	0.48	1.61	0.44	−7.14	0.20
School orientation	1.33	0.37	1.68	0.31	33.58	1.03
Warmth	0.83	0.19	0.90	0.14	12.22	0.42
Adolescent achievement						
Arithmetic reasoning	94.37	11.78	104.18	12.30	26.74	0.81
Word recognition	97.33	14.09	106.12	14.04	20.59	0.62
Reading comprehension	93.08	11.81	102.14	12.23	24.78	0.75

Note. N = 4,406. SES = socioeconomic status. All *t*s are significant at $p < .001$.

fixed to the square root of the internal consistency reliability of the indicator, and the residuals were fixed to the indicators' error variances (i.e., $1 - \alpha$) times their variances (Bollen, 1989).

SEM

The main analyses assessed the intergenerational model in Figure 1 with latent variable SEM. Although the path between ethnicity and adolescent achievement was predicted to be zero, it was allowed to freely vary to get a better estimate of the indirect and direct effects of ethnicity. The constrained model had an acceptable fit with the data, $\chi^2(143, N = 4{,}406) = 3{,}045$, CFI = .92, GFI = .93, RMSEA = .07. As predicted, the direct effects of ethnicity reduced to no difference between African Americans and European Americans. Also, most of the hypothesized paths in the model were significant in the expected direction. However, an

Table 3
Factor Loadings of Indicators for Latent Variables With Multiple Indicators

Multi-indicator latent variables	Unstandardized factor loadings	Standardized factor loadings	R^2
Grandparent SES			
Grandparent SES → Occupational prestige	1.00[a]	.62	.38
Grandparent SES → Education	729.95**	.68	.46
Grandparent SES → Literary resources	9.14**	.75	.56
Mother's achievement			
Mother's ach. → Arithmetic reasoning	1.00[a]	.93	.86
Mother's ach. → Word recognition	0.66**	.79	.62
Mother's ach. → Reading comprehension	0.41**	.88	.77
Family SES			
Family SES → Occupational prestige	1.00[a]	.72	.52
Family SES → Poverty status	278.27**	.44	.20
Family SES → Wealth	168,478.32**	.62	.38
Adolescent achievement			
Adolescent ach. → Arithmetic reasoning	1.00[a]	.92	.85
Adolescent ach. → Word recognition	0.85**	.77	.60
Adolescent ach. → Reading comprehension	1.10**	.89	.79

Note. N = 4,406. SES = socioeconomic status; ach. = achievement.
[a] Unstandardized factor loading was fixed to equal 1.00 and was not tested for significance. $\chi^2(48) = 1{,}040.13$, $p < .001$, CFI = .97, adjusted goodness-of-fit index = .96, root-mean-square error of approximation = .06.
** $p < .001$.

examination of the modification indices suggested that the model could be substantially improved by adding a path from grandparents' SES to family SES and a direct path from mother's achievement to adolescent's achievement. The modification indices also suggested adding direct paths from ethnicity to mother's achievement, family SES, and three of the parenting variables. This modified model is presented in Figure 2. The modified model resulted in a good fit to the data, $\chi^2(136) = 2{,}118$, $p < .001$, CFI = .95, AGFI = .95, RMSEA = .06, and a significant improvement in model fit compared with the conceptual model, $\Delta\chi^2(7) = 927$, $p < .001$. The variables in the model explained 57% of the variance in adolescent achievement scores.

The modified model shows that the large ethnic gap reduced to a significant but very small difference between African American and European American adolescents, with African American scores predicted to be higher by 1.6 points if everything else was equal (see Table 4). The analysis also shows that the effects of ethnicity were not completely explained by the originally proposed chain of intergenerational events. As suggested by the modification indices, ethnic differences in mothers' achievement, family SES, and three of the parenting variables could not be explained by the earlier variables in the hypothesized model. The model also suggests that grandparents' SES was mediated by their daughters' achievement and SES. Although most of the effects of mothers' achievement were mediated by their family's later SES, as predicted, mothers' achievement still had an important unique effect on their adolescents' achievement. Family SES had equally strong effects on each of the parenting variables, and as predicted in the original model, family SES was completely mediated by the parenting variables in the model. Four of the parenting variables also had significant direct effects on adolescent achievement. On average, the fewer arguments the parents and adolescents had about the rules, the more the family allowed adolescent decisions, and the more school-oriented the home environment, the better the adolescents tended to perform on the tests. A smaller but significant direct effect of household chores also implied that adolescents with high levels of household chore responsibilities had poorer achievement. Parental monitoring and warmth did not have a significant direct effect on adolescent achievement after the other variables were controlled.

Discussion

The primary purpose of the current study was to determine if various intergenerational family factors could account for the Black–White achievement gap in adolescence. Previous studies tended to use a smaller number of factors and/or relatively unreliable parenting measures and explained 40% to 50% of the ethnic achievement gap at best (e.g., Brooks-Gunn et al., 1996, 2003; Phillips et al., 1998). We predicted that by addressing their limitations with a larger set of reliable intergenerational SES and parenting factors, most of the ethnic differences in achievement would be explained. The results were very much in line with this prediction. The large, 10-point advantage European American adolescents have over African American adolescents was reduced to zero once the intergenerational family factors in the study were statistically controlled. This suggests that if their grandparents had

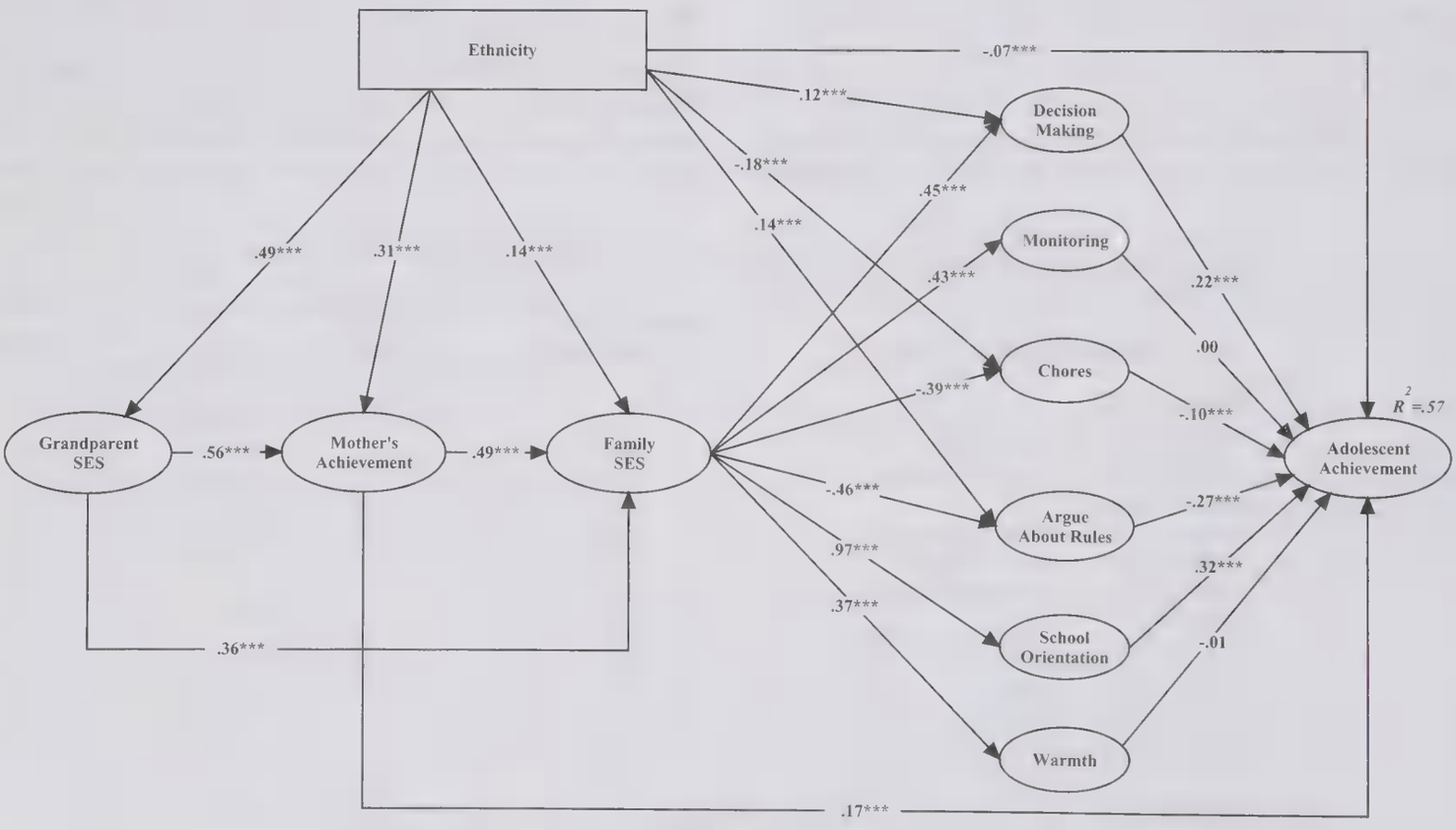


Figure 2. Modified conceptual model with direct and indirect effects. Standardized solution is shown. $\chi^2(136) = 2{,}118$, $p < .001$, comparative fit index = .95, adjusted goodness-of-fit index = .95, RMSEA = .06. $N = 4{,}406$ adolescents. SES = socioeconomic status. *** $p < .01$.

Table 4
Parameter Estimates and Significant Levels for Direct and Indirect Effects on Adolescent Achievement

Variable	Direct effects			Indirect effects		
	B	*SE*	β	*B*	*SE*	β
Ethnicity	−1.61***	0.41	−.07	10.61***	0.38	.45
Grandparent SES				24.63***	0.95	.45
Mother's achievement	0.26***	0.04	.17	0.41***	0.03	.28
Family SES				26.22***	1.65	.56
Decision making	15.56***	1.35	.22			
Parental monitoring	−0.05	0.37	−.00			
House chores	−5.00***	1.10	−.10			
Argue about rules	−6.98***	0.41	−.28			
School-oriented	9.74***	1.23	.32			
Warmth	−0.86	1.08	−.01			

Note. Standard errors and *p* values for the indirect effects were estimated with 500 bootstrap samples in AMOS 16.0. Ethnicity was coded as Black = 0 and White = 1. $N = 4{,}406$ adolescents. SES = socioeconomic status.
*** $p < .001$.

equal SES, their parents had the same achievement and SES, and they were exposed to the same parenting practices, the ethnic groups would score the same on achievement tests.

A further purpose of the study was to test a conceptual model of the process by which family factors have facilitated the development of ethnic achievement gaps over the past few generations. The model postulates that ethnic differences in SES make it more likely that European American parents will use parenting practices associated with achievement than African American parents. This difference in parenting facilitated ethnic differences in achievement. The model further postulates that this process is repeated in the following generations because those adolescents with higher achievement grow up to have better jobs and make more money. Their higher SES and experiences with their own parents increases the likelihood that they will use more academically oriented parenting with their own children. Thus, the model suggests that European Americans' SES advantages from prior generations help produce future generations' advantages in achievement, primarily through more achievement-oriented parenting.

Overall, we found strong support for a slightly modified version of the hypothesized intergenerational model. For instance, it was argued that one reason European American adolescents outperform African American adolescents is that their grandparents tended to have many more social and economic resources than African American grandparents. We found support for this assumption. The sizes of the economic and educational differences between European American and African American grandparents in this study were rather large. Furthermore, these past-generation advantages in occupational status, education, and literary resources predicted higher test scores for future generations. Thus, even though grandparents' resources were assessed in the late 1970s, they still had moderate to large relationships to their grandchildren's test scores in the mid- to late 1990s.

Also as expected, the effects of grandparents' resources were mediated by their own children's resources. We predicted that their daughters' achievement would account for all of the grandparents' SES, but this was not entirely the case. Grandparents' SES also had a direct effect on their grown daughters' family SES, above and beyond their daughters' achievement as adolescents. This suggests that besides higher achievement, much of the SES advantages enjoyed by European American parents may be due to inheritance and other direct forms of assistance (Gittleman & Wolff, 2002). This finding also supports the intergenerational transmission of resources theory, which argues that grandparents' primarily impact their grandchildren through their effects on their own children (Phillips et al., 1998). As a result, when parents pass down social and economic resources to their children, they are essentially passing much of those resources down to later generations of their family as well.

As found in several previous studies (e.g., Crane, 1996), mothers' achievement was highly correlated with adolescent achievement and most other factors in the study. One interesting finding was that the ethnic differences in mothers' test scores were much larger than the ethnic differences in their adolescents' test scores, even before any background factors were controlled. In fact, the Black–White test score gap decreased by 50% in this one generation. This finding could be an artifact of differences in the tests. However, the tests cover basically the same content. Also, because the gap in SES also decreased from the grandparents' generation to the parents' generation, and because these factors were correlated with achievement, it is logical to expect that a reduction in the test score gap between the parents' generation and their adolescents' generation would occur.

The intergenerational model further predicted that the effects of mothers' achievement on their adolescents' achievement would be mediated by family SES. This was partially supported. Most of the effects of mothers' achievement were explained by family SES and the parenting practices later in the chain of events, but mothers' achievement still had a direct effect on adolescent achievement. There are many possible reasons for this. The remaining direct effect could be the genetic contribution of mothers' innate intelligence, which cannot be explained by social factors. This perspective certainly has proponents (e.g., Herrnstein & Murray, 1994). However, no conclusions about genetic contributions can be drawn from these data. This is especially the case because the small remaining ethnic gap in adolescent achievement is in the opposite direction from what most supporters of the hereditarian perspective would have predicted. Regardless, it is more likely the case that unmeasured factors, such as the degree to which mothers read to their children or exposed them to the children of other educated parents, accounts for the remaining effects of mothers' achievement. One strong possibility is that more educated mothers have greater confidence in their ability to help their adolescents (McCarthey, 2000), and they know better how to navigate the educational environment by teaching their children how to take tests and being more actively involved at school (Hoover-Dempsey & Sandler, 1997). Future studies should examine other potential mediating factors of parents' own achievement on their children's achievement.

It was also expected that parents' economic resources would account for much of the ethnic achievement gap through their effect on parenting. This hypothesis was fully supported. Although the ethnic differences in SES were smaller than in the grandparents' generation, European American parents still had much higher levels of wealth and prestigious occupations and a lower likelihood of living in poverty than African American parents. Each of these factors also correlated strongly with adolescent achievement.

There was also a strong direct path between ethnicity and parents' SES, even after the other variables in the model were controlled. This implies that one reason European American children outperform African American adolescents is that their parents have higher levels of SES.

As predicted by the model and found in prior studies (Guo & Harris, 2000; Linver et al., 2002; Mistry et al., 2002), parents' SES was completely mediated by the parenting practices assessed in the study. The higher the parents' SES, the more likely they were to use achievement-oriented parenting practices. Specifically, the degree to which adolescents argued with parents was a strong predictor of achievement, net of all the other variables in the study. The higher adolescents were on that measure, the worse off they were in achievement. The reason for this is not clear. This variable could be a marker for lack of behavioral control, which is consistently related to poor achievement for all ethnic groups (Mandara, 2006; Spera, 2005). It could also just as easily be a marker of adolescents responding negatively to excessive behavioral control. Considering the strong relationship of this variable to achievement, future studies should explore the nature of this variable.

Another very important parenting predictor was the degree of freedom adolescents had to make certain decisions. The more decision-making freedom they had, the better adolescents performed on the tests. This may be because adolescents who are generally well behaved and who perform well in school receive rewards from parents, such as freedom to make choices. Thus, the direction of the arrow may be reversed in the real world. Other theorists have suggested that allowing adolescents' freedom to make developmentally appropriate decisions is an aspect of some parents' general philosophy about what is best for children (Baumrind, 1996). It may also be that decisions are associated with intellectual development because those regularly allowed autonomy of decisions get more experience at thinking about the pros and cons of various options (Mann, Harmoni, & Power, 1989). These everyday cognitive exercises may act as training for cognitive ability and achievement tests.

Similarly, one of the main parenting predictors of adolescent's academic success in this study was the school orientation of the home. In the current study, this included having an abundance of books at home, going on intellectually stimulating field trips to destinations such as museums, and taking music or other lessons. This is likely related to the reasons decision making was so strongly related to achievement. Constantly being exposed to intellectually stimulating material at home that mimics the school environment likely helps to teach adolescents the skills assessed in most achievement tests. It should also make learning novel concepts easier, because it facilitates more contextual background knowledge (Pazzani, 1991).

When parents did not burden adolescents with excessive amounts of household chores, the adolescents also tended to score higher on the tests, over and above the effects of the other parenting and SES variables. Although the effect was not very large, other studies have found similar results (Shanahan & Flaherty, 2001). The most likely reason is that excessive time spent on cooking, washing dishes, and cleaning the house takes time and energy away from schoolwork. Because the effect of household chores was not very large, a small to moderate amount of regular household responsibilities is likely not problematic. Future studies should examine possible curvilinear relationships with household chores and adolescent achievement.

African American and European American adolescents also perceived very different parenting experiences. They differed on all of the parenting practices. Even after controlling for grandparents' SES, mothers' achievement, and family SES, European American parents were still more likely to allow their adolescents freedom to make decisions and offer a more school oriented home environment than African American parents. African American parents also gave their adolescents more household chores and had more arguments with their adolescents about the rules than European American parents. The differences in the specific parenting practices explained much of the ethnic achievement gap. Because two generations of reliable SES and achievement variables were controlled, the ethnic differences in these parenting practices most likely represent different cultural beliefs about parenting (Mandara, 2006). Regardless of cultural beliefs, these results suggest that they are important reasons why African American adolescents do not perform as well on achievement tests as European American youth.

Summary and Implications

The overall pattern of results was clear. The ethnic achievement gap has been significantly reduced over the past few generations, but the social and economic differences of previous generations continue to advantage some and hinder others. Grandparent and parental social and economic resources explained the entire gap in achievement between African Americans and European Americans. This implies that African American parents who have high levels of formal education, wealth, prestigious occupations, and a school-oriented home environment, and use various parenting practices associated with achievement, tend to have adolescents who score as high as European American adolescents who come from similar backgrounds. Therefore, the general improvement in the quantity and quality of education, reduction in poverty, and more achievement oriented parenting in African American communities over the past generation may explain their relative increase in achievement (Ceci, 1991; Grissmer, Flanagan, & Williamson, 1998; McLoyd, 1998).

However, these results must be interpreted in light of some limitations. Although this study has many advantages over previous studies, including a relatively large sample and a diverse set of family factors, as with all correlational studies, causation is still a question. Even though many design features were implemented to reduce the overlap between the predictors and outcomes, much uncontrollable overlap remained. For instance, it is still possible that children's achievement influences parenting as much as parenting influences achievement. Parents may respond to poor achievement in the early grades by being more involved and increasing the educational experiences in the home.

Another problem with our ability to make causal claims is that many other potentially important factors were not assessed. For instance, exposure to family violence, parents' mental health, grandparents' parenting, and time spent with their grandparents may have important and unique effects on test scores and the achievement gap. We also had relatively crude measures of behavioral control and no information on parents' use of psychological control. These variables may shed more light on the factors

important to adolescent achievement and the achievement gap. Furthermore, a few of the variables included in the study had relatively low internal consistencies. A strong point of the study compared to prior achievement gap studies was the use of SEM to help adjust for measurement error. However, as the reliability of the indicators increases, latent variable modeling produces more accurate estimates (Bollen, 1989). Thus, these results can still be interpreted as somewhat conservative, and studies with more reliable indicators may result in slightly different estimates.

In spite of the limitations, these findings have many important implications for policy, practice, and educational research. One of the most important implications of this study is that it illustrates the importance of SEM for producing more powerful and accurate estimates (Bollen, 1989) of the effects of family factors. This study also shows how a more comprehensive set of measures can account for more of the achievement gap. For instance, a measure of parental warmth was a relatively weak predictor of achievement once the SES and other parenting factors were included in the model. A variation of this instrument has been the primary measure of parenting used in previous achievement gap studies (e.g., Brooks-Gunn et al., 1996, 2003). This study shows that parenting practices explain a great deal more about achievement than can be captured by just assessing parental warmth or even cognitive stimulation in the home.

Probably the most important implication of this study is that the entire gap in achievement between African American and European Americans may be due to modifiable social factors as opposed to immutable genetic factors. This suggests that researchers should not so easily assume that error variance in regression models is due to genetic effects (Block, 1995). This is important because it could help to transform social stereotypes about differences in innate intelligence and modify the negative and low expectations many teachers have of low-income and African American students (Ferguson, 2003; McKown & Weinstein, 2008; Rubie-Davies, Hattie, & Hamilton, 2006).

This study may also have important implications for the development of parent training prevention–interventions. Parenting explained most of the achievement gap, even after very good measures of SES and mothers' achievement were controlled. This implies that the parenting effects were as much a result of parental beliefs, customs, knowledge, and priorities about what practices are best for their adolescents as the economic class they come from (Mandara, 2006). Thus, it seems clear that one promising and practical way to help reduce the current achievement gap is through academically oriented parenting interventions (Mandara & Murray, 2007). Although African American and European American parents may have different cultural beliefs about parenting, the results of this study suggest that the achievement gap between their adolescents would be dramatically reduced if they used many of the parenting practices assessed in this study to the same degree.

The results of this study do not minimize the contribution of school, neighborhood, or peer factors typically studied in educational research. What this study shows is that family factors are major predictors of achievement and the achievement gap as well. Future research that combines comprehensive sets of well measured variables at all levels of student's social ecology will be best suited for uncovering the unique effects of each factor and the mechanisms by which these factors relate to achievement.

References

Arbuckle, J. L. (2007). *Amos 16.0 user's guide.* Spring House, PA: Amos Development.

Baumrind, D. (1996). The discipline controversy revisited. *Family Relations: Journal of Applied Family & Child Studies, 45,* 405–414.

Block, N. (1995). How heritability misleads about race. *Cognition, 56,* 99–128.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bradley, R. H., Caldwell, B. M., Rock, S. L., Barnard, K. E., Gray, C., Hammond, M. A., et al. (1989). Home environment and cognitive development in the first three years of life: A collaborative study involving six sites and three ethnic groups in North America. *Developmental Psychology, 25,* 217–235.

Bradley, R. H., Corwyn, R. F., Pipes McAdoo, H., & Garcia Coll, C. (2001). The home environments of children in the United States: Part I. Variations by age, ethnicity, and poverty status. *Child Development, 72,* 1844–1867.

Brody, G. H., & Flor, D. L. (1997). Maternal psychological functioning, family processes, and child adjustment in rural, single-parent, African American families. *Developmental Psychology, 33,* 1000–1011.

Brody, G. H., Stoneman, Z., & McCoy, J. K. (1994). Contributions of protective and risk factors to literacy and socioemotional competency in former Head Start children attending kindergarten. *Early Childhood Research Quarterly, 9,* 407–425.

Brooks-Gunn, J., Klebanov, P. K., & Duncan, G. J. (1996). Ethnic differences in children's intelligence test scores: Role of economic deprivation, home environment, and maternal characteristics. *Child Development, 67,* 396–408.

Brooks-Gunn, J., Klebanov, P. K., Smith, J., Duncan, G. J., & Lee, K. (2003). The Black–White test score gap in young children: Contributions of test and family characteristics. *Applied Developmental Science, 7,* 239–252.

Byrnes, J. P. (2003). Factors predictive of mathematics achievement in White, Black, and Hispanic 12th graders. *Journal of Educational Psychology, 95,* 316–326.

Caldwell, B., & Bradley, R. (1984). *Home Observation for Measurement of the Environment–Revised edition.* Little Rock: University of Arkansas, Little Rock.

Campbell, F. A., Pungello, E. P., Miller-Johnson, S., Burchinal, M., & Ramey, C. T. (2001). The development of cognitive and academic abilities: Growth curves from an early childhood educational experiment. *Developmental Psychology, 37,* 231–242.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology, 27,* 703–722.

Chase-Lansdale, P. L., Mott, F. L., Brooks-Gunn, J., & Phillips, D. A. (1991). Children of the National Longitudinal Survey of Youth: A unique research opportunity. *Developmental Psychology, 27,* 918–931.

Conger, R. D., Wallace, L. E., Sun, Y. M., Simons, R. L., McLoyd, V. C., & Brody, G. H. (2002). Economic pressure in African American families: A replication and extension of the family stress model. *Developmental Psychology, 38,* 179–193.

Crane, J. (1996). Effects of home environment, SES, and maternal test scores on mathematics achievement. *Journal of Educational Research, 89,* 305–314.

DeNavas-Walt, C., Cleveland, R. W., & Webster, B. H., Jr. (2003). *Income in the United States: 2002* (Consumer Population Reports, P60–221). Washington, DC: U.S. Government Printing Office.

Duncan, G. J., Brooks-Gunn, J., & Klebanov, P. K. (1994). Economic deprivation and early childhood development. *Child Development, 65,* 296–318.

Eller, T. J. (1994). *Household wealth and asset ownership: 1991* (Current

Population Reports, P70–34). Washington, DC: U.S. Government Printing Office.

Ferguson, R. F. (2003). Teachers' perceptions and expectations and the Black–White test score gap. *Urban Education, 38,* 460–507.

Fryer, R., & Levitt, S. (2004). Understanding the Black–White test score gap in the first two years of school. *Review of Economics and Statistics, 86,* 447–464.

Gittleman, M., & Wolff, E. N. (2002). Ethnic differences in patterns of wealth accumulation. *Journal of Human Resources, 39,* 193–227.

Goldstein, S. E., Davis-Kean, P. E., & Eccles, J. S. (2005). Parents, peers, and problem behavior: A longitudinal investigation of the impact of relationship perceptions and characteristics on the development of adolescent problem behavior. *Developmental Psychology, 41,* 401–413.

Grissmer, D., Flanagan, A., & Williamson, S. (1998). Why did the Black–White score gap narrow in the 1970s and 1980s? In C. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 182–226). Washington, DC: Brookings Institution Press.

Guo, G., & Harris, K. M. (2000). The mechanisms mediating the effects of poverty on children's intellectual development. *Demography, 37,* 431–447.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life.* New York: Free Press.

Hoover-Dempsey, K., & Sandler, H. (1997). Why do parents become involved in their children's education? *Review of Educational Research, 67,* 3–42.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Jencks, C., & Phillips, M. (1998). The Black–White test score gap: An introduction. In C. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 1–54). Washington, DC: Brookings Institution.

Kamo, Y. (2000). Ethnic differences in extended family households. *Sociological Perspectives, 43,* 211–229.

Linver, M. R., Brooks-Gunn, J., & Kohen, D. E. (2002). Family processes as pathways from income to young children's development. *Developmental Psychology, 38,* 719–734.

Mandara, J. (2006). How family functioning influences African American males' academic achievement: A review and clarification of the empirical literature. *Teachers College Record, 10,* 205–222.

Mandara, J., & Murray, C. B. (2007). How African American families can facilitate the academic achievement of their children: Implications for family-based interventions. In J. Jackson (Ed.), *Strengthening the educational pipeline for African Americans: Informing policy and practice* (pp. 165–186). Albany: State University of New York Press.

Mann, L., Harmoni, R., & Power, C. (1989). Adolescent decision-making: The development of competence. *Journal of Adolescence, 12,* 265–278.

McCarthey, S. (2000). Home–school connections: A review of the literature. *Journal of Educational Research, 93,* 145–153.

McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *Journal of School Psychology, 46,* 235–261.

McLoyd, V. C. (1998). Socioeconomic disadvantage and child development. *American Psychologist, 53,* 185–204.

Mistry, R. S., Vandewater, E. A., Huston, A. C., & McLoyd, V. C. (2002). Economic well-being and children's social adjustment: The role of family process in an ethnically diverse low-income sample. *Child Development, 73,* 935–951.

Moore, E. G. (1986). Family socialization and the IQ test performance of traditionally and transethnicly adopted Black children. *Developmental Psychology, 22,* 317–326.

Oakes, J. (2005). *Keeping track: How schools structure inequality* (2nd ed.). New Haven, CT: Yale University Press.

Orr, A. J. (2003). Black–White difference in achievement: The importance of wealth. *Sociology of Education, 76,* 281–304.

Pazzani, M. (1991). The influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 416–432.

Phillips, M., Brooks-Gunn, J., Duncan, G. J., Klebanov, P., & Crane, J. (1998). Family background, parenting practices, and the Black–White test score gap. In C. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 103–145). Washington, DC: Brookings Institution.

Phillips, M., Crouse, J., & Ralph, J. (1998). Does the Black–White test score gap widen after children enter school? In C. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 229–272). Washington, DC: Brookings Institution.

Rubie-Davies, C., Hattie, J., & Hamilton, R. (2006). Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Psychology, 76,* 429–444.

Saltaris, C., Serbin, L. A., Stack, D. M., Karp, J. A., Schwartzman, A. E., & Ledingham, J. E. (2004). Nurturing cognitive competence in preschoolers: A longitudinal study of intergenerational continuity and risk. *International Journal of Behavioral Development, 28,* 105–115.

Shanahan, M. J., & Flaherty, B. P. (2001). Dynamic patterns of time use in adolescence. *Child Development, 72,* 385–401.

Spera, C. (2005). A review of the relationship among parenting practices, parenting styles, and adolescent school achievement. *Educational Psychology Review, 17,* 125–146.

Steinberg, L., Lamborn, S. D., Darling, N., Mounts, N. S., & Dornbusch, S. M. (1994). Over-time changes in adjustment and competence among adolescents from authoritative, authoritarian, indulgent, and neglectful families. *Child Development, 65,* 754–770.

Steinberg, L., Mounts, N. S., Lamborn, S. D., & Dornbusch, S. M. (1991). Authoritative parenting and adolescent adjustment across varied ecological niches. *Journal of Research on Adolescence, 1,* 19–36.

Strage, A., & Brandt, T. S. (1999). Authoritative parenting and college students' academic adjustment and success. *Journal of Educational Psychology, 91,* 146–156.

Taylor, L. C., Hinton, I. D., & Wilson, M. N. (1995). Parental influences on academic performance in African American students. *Journal of Child & Family Studies, 4,* 293–302.

Received September 7, 2006
Revision received May 11, 2009
Accepted May 18, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 879–896

0022-0663/09/$12.00 DOI: 10.1037/a0015945

Age-Related Differences in Achievement Goal Differentiation

Mimi Bong
Korea University

Validity of the 2 × 2 achievement goal framework for school-aged children and adolescents was examined, using self-report responses from 1,196 Korean elementary and middle school students. Confirmatory factor analysis models hypothesizing 4 distinct achievement goal factors demonstrated the best fit in all age groups. Nevertheless, achievement goals of these young students were strongly correlated with each other, regardless of the goal definition or valence. The correlation became increasingly weaker with the increasing age of the respondents. Students in Grades 1–4 endorsed a mastery-approach goal most strongly, but those in Grades 5–9 endorsed a performance-approach goal. Performance-avoidance and mastery-avoidance goals received significantly lower average ratings than did the 2 approach goals in all age groups. Whereas both mastery-approach and performance-approach goals correlated positively with self-efficacy, strategy use, and performance in math, only the performance-approach goal correlated positively with anxiety. Anxiety also correlated positively with the 2 avoidance goals. A performance-avoidance goal further demonstrated positive correlation with help-seeking avoidance, whereas a mastery-avoidance goal did so with strategy use.

Keywords: achievement goals, age differences, adolescence, mastery avoidance

Students demonstrate achievement behaviors for many different reasons. For some, it is the belief that acquiring new knowledge and mastering new skills will improve their competence that leads them to invest genuine effort in learning. Others study hard with the goal of outperforming their peers, because they believe doing so is the surest way to verify their superior ability. For yet others, the primary purpose of engaging in schoolwork is to neither improve their competence nor document their superiority but, rather, to hide their inadequacy from their teachers and peers. Achievement goals refer to these underlying reasons and purposes that explain why individuals demonstrate achievement-related behaviors in specific settings the way they do (Ames, 1992).

Research on achievement goals has risen as one of the most active areas in classroom motivation research during the past 15 years (Pintrich, 2003). Early contributors in this area conceptualized students' achievement goals within a dichotomous and unidimensional framework. Students' orientations toward learning and understanding, developing new skills, and mastering challenging tasks for the purpose of improving their competence were variously termed as *learning goals* (Dweck & Leggett, 1988), *mastery goals* (Ames, 1992), or *task-involvement* (Nicholls, 1984). These goals were thought to represent an adaptive end of the motivational continuum. In contrast, students' desires to outperform their peers and publicly validate their intellectual superiority, called *performance goals* (Ames, 1992; Dweck & Leggett, 1988) or *ego-involvement* (Nicholls, 1984), were viewed to represent a maladaptive end.

Whereas the adaptive nature of a mastery goal was arguably well documented within this framework, the presumed maladaptive nature of a performance goal was not. Across studies, students' performance goals demonstrated negative, nonsignificant, or even positive relationships with beneficial student outcomes, such as grades. This led several researchers to call for a distinction between the approach and the avoidance properties of a performance goal (Elliot & Harackiewicz, 1996; Middleton & Midgley, 1997; Skaalvik, 1997; Urdan, 2004). Subsequent studies provided empirical support for the proposed separation. A performance-approach goal demonstrates non-negative and, more often, positive associations with students' self-efficacy and academic performance. A performance-avoidance goal, on the contrary, displays negative associations with those adaptive outcomes but, instead, demonstrates positive associations with the maladaptive outcomes such as anxiety and use of self-defeating strategies (Elliot & Church, 1997; Middleton & Midgley, 1997; Skaalvik, 1997). This trichotomous framework is the most widely accepted view in contemporary achievement goal literature.

2 × 2 Achievement Goal Framework

More recently, researchers such as Elliot (1999) and Pintrich (2000) argued that the reasons why students would engage in particular academic pursuits could be better understood by simultaneously considering both their general purposes of engagement (i.e., goal valence) and the criteria they use to judge their own

Mimi Bong, Department of Education and Brain and Motivation Research Institute, Korea University, Seoul, Korea.

This research was inspired by a discussion with the late Paul R. Pintrich of the University of Michigan. I gratefully acknowledge his valuable insights, openness to new ideas, and willingness to collaborate with others with conflicting views. This work was supported by the 2007 Korea University Research Grant.

Correspondence concerning this article should be addressed to Mimi Bong, Department of Education, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-701, Korea. E-mail: mimibong@korea.ac.kr

performance (i.e., goal definition). According to this conceptualization, individuals construe competence as either a positive (i.e., success) or a negative possibility (i.e., failure). Positive possibilities induce an approach tendency, whereas negative possibilities invoke an avoidance tendency. Further, individuals could rely on either absolute, intraindividual criteria or normative, interindividual criteria when evaluating their own competence.

When applying this logic to classroom situations, students may demonstrate an approach orientation toward success by striving to master the new materials, while focusing on individual progress to assess the quality of their performance (i.e., mastery-approach goals). Some students may likewise exhibit an orientation to approach success but nonetheless gauge their competence by relative superiority of their performance to that of their peers (i.e., performance-approach goals). Alternatively, students could demonstrate an avoidance orientation and try to escape failure by concealing their relative incompetence in front of others (i.e., performance-avoidance goals). Still others may define incompetence as not performing as well as before or the best they could and try to circumvent failure by avoiding such possibilities (i.e., mastery-avoidance goals).

Elliot and McGregor (2001) presented the first piece of empirical evidence demonstrating independence of these four achievement goals in the 2 × 2 framework. Using exploratory and confirmatory factor analyses across three studies with responses from U.S. undergraduate students, the investigators demonstrated that the four hypothesized goal factors were clearly defined and separable from each other. Further, the four achievement goals showed different patterns of associations with a set of presumed antecedents and consequents. Overall need for achievement predicted both types of approach goals. Workmastery, self-determination, and perceived classroom engagement predicted a mastery-approach goal, whereas fear of failure, competitiveness, and SAT scores predicted a performance-approach goal. Fear of failure and self-determination were, respectively, common positive and negative predictors of the two avoidance goals. SAT scores negatively predicted a performance-avoidance goal, whereas classroom engagement positively predicted a mastery-avoidance goal.

Several researchers partially replicated Elliot and McGregor's (2001) original findings, such as empirical independence of the four achievement goals; positive correlation of a mastery-avoidance goal with performance-avoidance and mastery-approach goals; and positive correlation of performance-approach, performance-avoidance, and mastery-avoidance goals with fear of failure (e.g., Conroy, Elliot, & Hofer, 2003; Finney, Pieper, & Barron, 2004; Pastor, Barron, Miller, & Davis, 2004). Nevertheless, an empirical base for the 2 × 2 achievement goal framework is still in its nascent stage and requires a great deal more evidence before tenability of this framework can be assessed reliably. In particular, more evidence is needed regarding the applicability of the 2 × 2 framework for describing young children's achievement goals and the unique role of each achievement goal in individuals' achievement strivings.

Present Study

The primary objective of the present research was to test validity of the 2 × 2 achievement goal framework for Korean elementary and middle school students. It was of particular interest to replicate Elliot and McGregor's (2001) findings from their confirmatory factor analyses and see if the four hypothesized achievement goal latent variables clearly emerged from the responses of much younger, school-aged children and adolescents. Children at nine different grade levels participated in this study, which made it possible to detect potential age-related differences in the degree and the pattern of achievement goal differentiation. Having a group of Korean students as participants, the present study further provided a valuable opportunity to assess generalizability of the 2 × 2 framework across not only school levels but also cultures.

As Elliot and McGregor (2001) exemplified, a stringent test of construct validity calls for evidence of discriminant validity as well as explanatory utility for a meaningful and distinctive set of outcomes. Once the most defensible factor structure for the present sample was determined, it was thus of equal interest to examine the relationships of the achievement goal factors that emerged with other motivation and performance variables. In addition to these general validity concerns, there were issues more specific to the proposed achievement goal framework.

Does the 2 × 2 Framework Adequately Represent Young Children's Achievement Goals?

The majority of existing studies supporting the 2 × 2 framework relied on college samples. College students by nature are those who self-select themselves into an academic environment. Their achievement history, perceptions of competence, need for achievement or social approval, manners in which they respond to performance opportunities, and types of attributions they make for their successes and failures might be different and presumably less variable compared with those of the general population (Pastor et al., 2004). More pertinent to the present research, the goals and purposes college students embrace in achievement situations might not be the same with those recognized by younger students in similar situations. The first question posed in this study, therefore, concerned the meaningfulness of the goal valence and the goal definition distinctions in the 2 × 2 framework for school-aged children.

Developmental research generally shows that younger children's belief systems are less clearly differentiated than are those of older children. Although children as young as kindergarteners and first graders appear capable of distinguishing between different aspects of the belief system or their beliefs across different activity domains (Eccles, Wigfield, Harold, & Blumenfeld, 1993), the degree of such differentiation typically increases with age (Marsh, Craven, & Debus, 1991). A previous study with Korean secondary school students reported results consistent with this developmental trend by showing that the achievement goals of middle school students were much more strongly correlated with one another than were those of high school students assessed with the same survey (Bong, 2001). Similar results were observed between elementary school children and college undergraduates (Ross, Shannon, Salisbury-Glennon, & Guarino, 2002). In view of these findings, younger students' achievement goals were hypothesized to be less clearly differentiated than those of older students.

Nevertheless, young children might perceive larger discrepancy between the situations in which they primarily strive to achieve success and those in which they try to avoid failure. As a conse-

quence, there might exist noticeable division between their approach and avoidance achievement goals. Some of the potent antecedents of achievement goals include perceived competence, personality dispositions, and salient contextual features (A. J. Elliot, 2005; Elliot & Church, 1997; Elliot & Harackiewicz, 1994; Jagacinski, Madden, & Reider, 2001; Linnenbrink, 2005; Senko & Harackiewicz, 2002). Among these, perceived competence formed on the basis of past achievement history appears to be the most influential precursor that leads individuals down either an approach or an avoidance path (Brophy, 2005; Elliot, 2005). To construe competence as a negative possibility, individuals need to have experienced some failure in similar situations. However, young children might not have undergone sufficient failures just yet to perceive achievement situations as aversive stimuli. Moreover, they tend to discount the significance of failure, even when they experience one. A review of the developmental literature on children's definitions and criteria for assessing competence discovered that young children regard success as evidence of their competence but do not necessarily consider failure to be evidence of their incompetence (Stipek & Mac Iver, 1989).

Self-theories present another theoretical ground for the proposed discrimination of approach and avoidance achievement goals by young children. According to Dweck (2000), self-theories individuals hold regarding whether their personal qualities are fixed or malleable direct them into different behaviors and outcomes. In particular, implicit theory of intelligence or beliefs about malleability of one's intelligence determines how much individuals value competence acquisition versus competence validation. Subscribers of an "entity" theory of intelligence believe intelligence is something one either possesses or does not possess and hence deem it important to have their competence validated. In contrast, subscribers of an "incremental" theory of intelligence believe intelligence can be improved by investing effort. For these individuals, acquiring new competence would be far more important than having their current levels of competence validated. An entity theory is a logical predictor of performance-oriented achievement goals, as is an incremental theory of mastery-oriented achievement goals.

Most young children are incremental theorists, who believe in malleable and developing nature of ability (Dweck & Leggett, 1988). Failure for them simply means that they need more practice to become more competent (Stipek & Mac Iver, 1989). Because they seldom engage in achievement-related strivings for the purpose of avoiding impending failure, characteristics of the few unusual situations in which they find themselves struggling to avoid failure are likely accentuated in their minds. Moreover, the psychological gap children feel between the approach and avoidance situations should be greater than that between the competence acquisition and competence validation situations because, regardless of whether they ultimately seek task mastery or relative superiority, the latter two "approach" scenarios share positive valence.

By the time children advance to upper elementary school grades, they develop what Nicholls (1984) termed as differentiated conceptions of ability. Children now begin to infer higher ability when the same level of performance is obtained with less effort. Differentiated conceptions of ability do not develop until about the age of 10 (Stipek & Mac Iver, 1989), which coincides with the period that children start incorporating social comparative information for the purpose of evaluating competence (Ruble, Boggiano, Feldman, & Loebl, 1980). Compared with younger children, who subscribe to an incremental theory of intelligence and assess their competence by making intraindividual comparisons of their own past and present skill levels, older children gradually shift to interindividual comparisons for appraising the quality of their own performance. Both the goal valence and the goal definition dimensions would be equally central for these early adolescents, who not only take their failure experiences into account but also resort to normative criteria when gauging competence. Therefore, the four achievement goals proffered by the 2 × 2 framework are hypothesized to be more or less clearly defined among the older students along the lines of both the valence of achievement strivings and the definition of competence.

Are There Age-Related Differences in the Degree Children Endorse Each Achievement Goal?

Assuming that multiple achievement goals emerged from students' responses, the next question would be whether there were differences in the strengths of these goals by age. It was argued earlier that young children would judge achievement situations as largely appetitive and demonstrate approach, rather than avoidance orientations, owing to the salience of their own past success experiences and their beliefs in the incremental theory of intelligence. Research indeed demonstrates that young children are heavily oriented toward what appears to be mastery-approach goals. When Harter (1975) asked 4- and 10-year old children to play a cognitively challenging game, the 10-year-olds demonstrated strong mastery motivation such that they played the game for "the gratification inherent in discovering the solution" (p. 376). The 4-year-olds played the game primarily because it was fun and enjoyable, which was also a form of mastery motivation. Together, it seems quite reasonable to expect high levels of mastery-approach goals among young children.

A performance goal, by definition, requires the desire and the capability of individuals to engage in social comparison in a systematic fashion. As discussed above, very young children may lack the capability or, more important, the desire to actively pursue and use social comparative information for the purpose of judging their own competence. Children as young as kindergarteners and first graders show some interest in seeking information about how similar others perform the same task (France-Kaatrude & Smith, 1985; Ruble, Feldman, & Boggiano, 1976). However, Ruble and her colleagues discovered that the self-rated competence of the kindergarteners and the second graders in their experiments was not affected by this information. Only the self-ratings of the fourth graders showed consistent effects of the social comparison information provided (Ruble et al., 1980). When Butler (1989) assigned 5-, 7-, and 10-year-olds to competitive and noncompetitive situations, a majority of the 5-year-olds provided mastery-oriented reasons for glancing at their peers' work, even under competition. In contrast, most 10-year-olds observed their peers' work for relative ability assessment, even when they were in a noncompetitive condition.

These findings again corroborate Nicholls's (1984) claim regarding the onset of differentiated conceptions of ability. They also support Harter's (1990) argument that the dimensions taking on

particular significance for self-evaluation change across developmental stages. She proposed that the most salient content of self-representations during early to middle childhood involves temporal comparisons with own past performance, which changes to comparative assessments with peers during middle to late childhood (Harter, 1998). This proposal suggests a stronger mastery-approach goal among children in lower elementary school grades and stronger performance-oriented goals among those in upper elementary school grades and above. It also adds further credence to the earlier claim that the definition of competence, one of the two axes comprising the 2 × 2 framework, may not be too meaningful for younger children.

Ruble et al. (1980) offered several explanations regarding why the younger children in their experiments might have assigned less weight to the social comparison information. According to their conjecture, younger children most likely concentrated on the tasks that they were directly experiencing with than on the performance of others. Younger children also more likely aimed at improving their skills than evaluating their current levels of competence, because of their beliefs in the incremental nature of ability (Dweck & Leggett, 1988). Last, the environments to which younger children had been exposed put much less emphasis on social comparative information than did the typical learning environments to which older children were more accustomed.

Recent research on the characteristics of learning environments and their impact on students' achievement goals provides strong support for this last conjecture. Starting with Ames and Archer's (1988) seminal investigation, numerous studies showed that students express stronger mastery-approach goals and lower avoidance goals in a mastery-oriented environment in which task mastery and individual progress are emphasized. When perceiving an emphasis instead on grades and relative ability in their environment, students tend to pursue performance-oriented achievement goals (Church, Elliot, & Gable, 2001; Midgley, Anderman, & Hicks, 1995; Roeser, Midgley, & Urdan, 1996; Turner et al., 2002; Wolters, 2004). Moreover, learning environment is heavily mastery-oriented in lower elementary school grades and gradually becomes performance-oriented in upper elementary school grades (Stipek & Mac Iver, 1989). By the time students enter middle schools, their learning environment is typically heavily performance-oriented (Eccles et al., 1993). Consequently, students perceive a stronger emphasis on social comparison and evaluation in their middle school environment than in their elementary school settings (Anderman & Midgley, 1997; Harter, Whitesell, Kowalski, 1992; Midgley et al., 1995; Urdan & Midgley, 2003).

Therefore, along with the greater propensity of young children to rely on intraindividual comparisons when evaluating competence (Stipek & Mac Iver, 1989), available evidence quite unanimously suggests a stronger mastery-approach goal among the children in the lower elementary school grades. Advent of differentiated conceptions of ability and increasingly abundant social comparative cues in the learning environment renders stronger performance-approach and performance-avoidance goals more likely among the middle school students in the current sample. This latter hypothesis is tentative because several investigators witnessed that students did not necessarily increase their personal performance goals, even when their perceptions of the competition and relative ability focus in their learning environments significantly increased (Anderman & Midgley, 1997; Bong, 2005; Urdan & Midgley, 2003). No specific hypothesis was generated regarding age-related differences in a mastery-avoidance goal.

Is a Mastery-Avoidance Goal Part of Young Children's Reasons for Achievement Behaviors?

It seems premature to generate specific hypotheses regarding the mastery-avoidance goal at this stage, because its very existence is still being debated. Despite the fact that several studies have presented empirical evidence for the 2 × 2 framework and, by extension, a mastery-avoidance goal, a number of problems remain to be resolved.

The first such problem has to do with the conceptual definition. Elliot (2005) defines a mastery-avoidance goal as the "striving to avoid losing one's skills and abilities (or having their development stagnate), forgetting what one has learned, misunderstanding the material, or leaving a task incomplete" (p. 61). This definition does not make it too obvious how a mastery-avoidance goal is different from a perfectionist orientation. Further, it is not exactly in sync with that of a mastery-approach goal, which has the same goal definition but different valence and represents the striving for the purpose of developing one's competence. To make the definition of a mastery-avoidance goal parallel to that of its counterpart, a stronger focus should be placed on avoiding the prospect of not learning or not improving as much as possible than on avoiding the possibility of not performing the best one could with the knowledge and skills currently available.

It is also questionable how much the goal of preventing potential deterioration of current knowledge and skills is psychologically relevant for young children, who are still very much in the process of acquiring new knowledge and developing new skills. Elliot (2005) acknowledged that, in general, a mastery-avoidance goal might be less ubiquitous than the other three achievement goals. He contended that a mastery-avoidance goal might nonetheless be "quite common in some instances and for some types of individuals" (p. 61). His examples include perfectionists, the elderly, and individuals who feel that they have reached their peak and start focusing on not doing worse than their past performances. Exploring whether a mastery-avoidance goal forms an independent factor for school-aged children and adolescents is an important contribution of the present study.

The second problem has to do with the relationships of a mastery-avoidance goal with other constructs. With few exceptions, antecedents of a mastery-avoidance goal largely overlap with those of a performance-avoidance goal (Elliot & McGregor, 2001). It is fully expected that these two goals would share antecedents, because both of them are avoidance goals. However, some of the common antecedents are not consistent with theory. As introduced above, Dweck (Dweck & Leggett, 1988) suggested that incremental theorists who believe in the malleable nature of ability pursue competence acquisition over competence validation, unlike entity theorists, who consider ability to be fixed and proving one's competence to be more important. The individuals in the former category define competence as improvement from their own past capabilities, whereas those in the latter category do so as relative superiority of their capabilities to those of others. This is why an incremental theory of intelligence is viewed as a plausible antecedent of mastery-approach and mastery-avoidance goals, the same way an entity

theory is of performance-approach and performance-avoidance goals.

However, Elliot and McGregor (2001) found that it was an entity theory, instead of an incremental theory, that predicted the adoption of a mastery-avoidance goal. Also contradictory to the reasoning based on the implicit theory of intelligence and the definition of competence espoused in each achievement goal, the subscription of an incremental theory of intelligence negatively predicted individuals' mastery-avoidance goals. These results thus raise serious doubts about the utility of the goal definition dimension for setting apart the two avoidance goals and suggest that they might be a lot more similar in content than theoretically warranted.

Another conundrum entails the pattern of intercorrelation among the achievement goals. Theoretically, a mastery-avoidance goal should demonstrate positive correlation with a mastery-approach goal, as well as a performance-avoidance goal, because it shares the definition of competence with the former and the valence of achievement striving with the latter. These predictions were supported in the validation studies cited earlier. However, a mastery-avoidance goal also frequently demonstrated a significant positive relationship with a performance-approach goal (Conroy et al., 2003; Elliot & McGregor, 2001, Study 1; Finney et al., 2004). This finding is unexpected and difficult to explain because these two goals share neither the goal definition nor the goal valence.

Also of particular relevance to the present issue was Elliot and McGregor's (2001) observation that the four achievement goals in the 2 × 2 framework were linked differently to the use of study strategies, anxiety, and academic performance of their college student participants. Specifically, a mastery-approach goal predicted deep processing of the course material, whereas a performance-approach goal predicted better exam performance. A performance-avoidance goal predicted surface processing of the course material, disorganization during exam preparation, state test anxiety, worry, emotionality, and poorer exam performance. A mastery-avoidance goal similarly predicted disorganization, worry, and emotionality but not poorer exam performance. The investigators concluded that the paths associated with the mastery-avoidance goal were, on the whole, more maladaptive in nature than those related to the mastery-approach goal but more adaptive than those associated with the performance-avoidance goal.

One of the objectives of the present study, therefore, was to determine whether a mastery-avoidance goal of younger students, if successfully identified, would demonstrate such a distinctive pattern of relationships with other variables as reported in Elliot and McGregor (2001). Given the suspected conceptual and empirical overlap between the mastery-avoidance and performance-avoidance goals, it appears to be a worthwhile endeavor to document the unique relationships, if any, of the mastery-avoidance goal with diverse motivation, strategy use, and performance variables for young children and adolescents.

Method

Participants and Procedures

Six hundred eighty-four elementary school students from 18 classrooms at four different schools and 512 middle school students from 12 classrooms at two different schools located in Seoul and Kyung-gi Province, Korea, participated in this research. The data came from a larger research project on Korean students' motivation and learning. Approximately half of the elementary school participants and all middle school participants completed additional surveys on other variables of interest, which were not part of this study. Korean elementary and middle schools offer 6 and 3 years of schooling, respectively.

After visually inspecting the database and checking frequencies, I excluded from the sample students who did not have achievement data (n = 47) and who failed to answer a substantial portion of the survey (n = 2). The final breakdown of the elementary school sample was 104 first graders, 110 second graders, 99 third graders, 107 fourth graders, 110 fifth graders, and 117 sixth graders. The middle school sample comprised 137 freshmen, 242 juniors, and 121 seniors (i.e., roughly equivalent to U.S. Grades 7, 8, and 9, respectively). Gender distribution fluctuated somewhat across the grade levels (i.e., 44.4%–51.9% girls in the elementary school sample; 48.9%–53.7% girls in the middle school sample).

To help examine developmental trends and make the sample sizes suitable for large-sample techniques such as confirmatory factor analyses (CFA), I merged data from two adjacent elementary school grade levels to form the following four age groups: lower elementary school grades (Grades 1 and 2; n = 214), middle elementary school grades (Grades 3 and 4; n = 206), upper elementary school grades (Grades 5 and 6; n = 227), and middle school grades (n = 500). Missing responses were randomly distributed, ranging up to 9 cases per variable (8.2%) in the lower elementary school grades, and were less than 5 cases per variable (5.0%) in all other samples. Missing values were imputed with a series mean of each variable for each grade level.

Data collection took place in May 2000 during regular classroom hours. Elementary school teachers read each item on the survey to their students. Middle school students filled out the surveys independently. Students were told that no one outside the research team, including their teachers and parents, would have access to their individual responses. As a way of further ensuring confidentiality of responses, students sealed their completed surveys with the stickers provided by the research team before turning them in to their teachers. Teachers collected the surveys in an envelope and sealed it before leaving the classroom.

Measures

Because of time constraints and anticipated differences in reading speed and fatigue across grade levels, surveys for younger students consisted of a considerably fewer number of variables compared with those for older students. Whereas achievement goals and self-efficacy were assessed at all grade levels, help-seeking avoidance was assessed from Grade 3 and up, and anxiety was assessed from Grade 5 and up. Use of cognitive and self-regulatory strategies was assessed in only the middle school sample. The grade levels providing data for each variable are indicated in parentheses below. All survey items referred to students' math class or math as a subject matter area. Students responded to each item on a 1 (*not at all true*) to 5 (*very true*) response scale.

All achievement goal and self-efficacy survey items, except for the mastery-avoidance goal, were first translated into Korean by Mimi Bong. A second coder independently translated these items into Korean. A third coder confirmed that the two translated

versions were consistent with each other, as well as the original version in English in content. Mimi Bong also translated all other items used in this study. The translated items on achievement goals and self-efficacy (Bong, 2001, 2005, 2008; Joo, Bong, & Choi, 2000), help-seeking avoidance (Bong, 2008), and cognitive and self-regulatory strategy use (Bong, 2008; Joo et al., 2000) had been used successfully in prior research with different groups of Korean students of varying ages.

Achievement goals (all grades). The original Mastery (i.e., mastery-approach; six items), Performance-Approach (five items), and Performance-Avoid (six items) Goal Orientation scales of the Patterns of Adaptive Learning Scales (Midgley et al., 2000) were adopted. Six mastery-avoidance goal items were developed for the present study. The Appendix presents descriptions and factor loadings of all 23 achievement goal items used in the present study.

Math self-efficacy (all grades). The Self-Efficacy scale of the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich & De Groot, 1990) was used. Three normative comparison items on the original scale were excluded for theoretical reasons (see, e.g., Bong & Clark, 1999; Bong & Skaalvik, 2003). The remaining items asked about students' subjective convictions for successfully carrying out various tasks in math class (six items; e.g., "I am sure that I can do an excellent job on the problems and tasks assigned for math class").

Help-seeking avoidance (elementary school grades 3–6; all middle school grades). Five items were adopted from Ryan and Pintrich (1997) and Ryan, Gheen, and Midgley (1998). These items assessed students' tendency to avoid seeking help in math class, even when it was needed (e.g., "When I don't understand my math work, I often guess instead of asking someone for help").

Anxiety (elementary school grades 5–6; all middle school grades). Three items from the Anxiety scale of the MSLQ were used. These items concerned worry and cognitive interferences during tests (e.g., "I am so nervous during a math test that I cannot remember things I have learned").

Cognitive strategy use (all middle school grades). Six items from the Cognitive Strategy Use scale of the MSLQ, pertaining to the use of rehearsal, elaboration, and organizational strategies were included (e.g., "When I do my math homework, I try to remember what the teacher said in class so I can answer the questions correctly").

Self-regulatory strategy use (all middle school grades). Six items on metacognitive and effort management strategy use were adopted from the Self-Regulation scale of the MSLQ (e.g., "When I study math, I ask myself questions to make sure I know the material I have been studying").

Math performance (all grades). First graders in Korean elementary schools receive neither grades nor progress reports. Elementary school students in Grades 2 to 6 receive end-of-semester progress reports made up of written comments from the teacher. Middle school students receive grades in each academic subject. For the present study, teachers at participating elementary schools were asked to provide ratings of their students' math performance on a scale of 1 (*below average*), 2 (*average*), and 3 (*above average*), on the basis of the data they collected in preparation for the progress reports. For the middle school students, end-of-semester math final examination scores were collected from the school databases with the permission of the school principals. Possible scores on these exams ranged between 0 and 100.

Overview of Data-Analysis Strategies

Because the primary purpose of the present study was to test validity of the 2 × 2 achievement goal framework for younger students, we specified a number of CFA models and compared them within each age group. Repeating Elliot and McGregor's (2001) procedures, we first tested a CFA model with four a priori achievement goal factors, followed by models that merged achievement goals of the same definition, the same valence, or both. This set of procedures addressed the questions of age-related differences in achievement goal differentiation and empirical validity of a mastery-avoidance goal among school-aged children and adolescents.

The question of age-related variations in the strengths of achievement goal endorsement was first examined by a multivariate analysis of variance (MANOVA) with age group and gender as between-subjects factors. Dependent variables were the set of achievement goals judged to best represent the data in the preceding CFAs. Significant differences detected in the MANOVA were further probed by univariate analyses and, when applicable, post hoc procedures. Within each age group, I performed multiple paired samples t tests for all possible pairs of achievement goals (e.g., mastery-approach vs. performance-approach goals) to test whether students in each age group endorsed a particular achievement goal more strongly than they did others. To maintain an experiment-wise $\alpha_E < .05$, I set the a priori alpha level for all post hoc analyses and paired-samples t tests at .001 (Hinkle, Wiersma, & Jurs, 2003; Stevens, 1992).

Finally, I specified a full CFA model with all variables for each age group to examine the pattern of relationships of the achievement goal factors with self-efficacy, help-seeking avoidance, anxiety, cognitive and self-regulatory strategy use, and performance in math.

Results

Table 1 reports descriptive statistics and reliability of the scales. Two developmental trends are obvious. Responses of the younger students were less reliable than those of the older students. With few exceptions, responses of the younger students were also higher than those of the older students. Mean ratings of a mastery-approach goal, a performance-approach goal, and math self-efficacy were especially high in the two younger age groups, Grades 1–4.

Tables 2 and 3 present zero-order correlation coefficients among the variables within each age group. Several findings are noteworthy. First, correlation coefficients among the four achievement goals were generally larger in the two younger samples. In particular, performance-approach and performance-avoidance goals ($r = .58$) and performance-avoidance and mastery-avoidance goals ($r = .69$) were strongly correlated in the lower elementary school sample. Second, performance-approach and mastery-avoidance goals demonstrated significant and substantial positive correlation in all age groups ($.22 \leq rs \leq .45$). Third, whereas none of the achievement goals correlated significantly with math performance in the lower elementary school sample, a mastery-approach goal

Table 1
Descriptive Statistics and Reliability of Scales

Scale	Lower elementary: Grades 1–2 (n = 214)			Middle elementary: Grades 3–4 (n = 206)			Upper elementary: Grades 5–6 (n = 227)			Middle school: Grades 7–9 (n = 500)		
	M	*SD*	α	*M*	*SD*	α	*M*	*SD*	α	*M*	*SD*	α
Mastery-approach goal	4.40	0.62	.66	4.00	0.73	.70	3.46	0.82	.78	3.22	0.92	.84
Performance-approach goal	4.07	0.80	.60	4.00	0.76	.61	3.75	0.76	.70	3.51	0.81	.69
Performance-avoidance goal	3.70	1.06	.77	3.06	1.07	.79	2.58	0.81	.74	2.41	0.78	.78
Mastery-avoidance goal	3.51	0.96	.70	3.39	1.02	.79	2.96	0.90	.79	2.99	0.81	.74
Math self-efficacy	4.22	0.70	.67	4.03	0.81	.83	3.57	0.88	.87	3.36	0.92	.88
Help-seeking avoidance	—	—	—	2.32	0.99	.71	2.36	0.89	.79	2.50	0.80	.74
Anxiety	—	—	—	—	—	—	2.86	0.92	.57	3.04	0.98	.63
Cognitive strategy use	—	—	—	—	—	—	—	—	—	3.35	0.73	.74
Self-regulatory strategy use	—	—	—	—	—	—	—	—	—	3.19	0.66	.60
Math performance	2.39	0.70	—	2.30	0.72	—	2.24	0.80	—	70.24	22.56	—

Note. Dashes indicate variables not assessed.

correlated positively and a performance-avoidance goal correlated negatively with math performance in the middle elementary (i.e., Grades 3 and 4) and older samples. A performance-approach goal demonstrated significant positive correlation with math performance in the upper elementary (i.e., Grades 5 and 6) and middle school samples (i.e., Grades 7–9), but it was always in smaller magnitude than that demonstrated by a mastery-approach goal. Fourth, math self-efficacy and cognitive and self-regulatory strategy use, when assessed, demonstrated positive correlation with math performance, whereas help-seeking avoidance and anxiety demonstrated negative correlation. Of all the variables assessed, math self-efficacy displayed the strongest correlation with math performance in all age groups.

Testing Age-Related Differences in Achievement Goal Differentiation

A total of eight CFA models with a different number of achievement goal latent variables were fitted separately in each age group, starting with the four-factor (i.e., 2 × 2) model. Seven of these models were direct replication of those tested by Elliot and McGregor (2001). A unidimensional model with a single achievement goal factor was also tested in this study. Scores on each survey item functioned as indicators, which I centered around item means to avoid potential problems such as instability in parameter estimates that might be caused by strong collinearity among the variables. All CFAs were performed with the EQS program (Bentler, 1995). Goodness-of-fit indexes such as Bentler-Bonett non-normed fit index (NNFI), comparative fit index (CFI), average absolute standardized residuals (res.), and statistical significance of factor loadings and factor variances were considered in evaluating the model fit. Values of NNFI and CFI greater than .90 and average residuals less than .10 generally represent acceptable model fit (Kline, 1998).

Initial 2 × 2 models with all 23 items loading on their hypothesized factors were associated with unsatisfactory fit in all age groups with values of NNFI and CFI ranging between .80 and .85. Although all factor loadings, factor variances, and error variances were statistically significant at $p < .05$, several items had loadings not substantial enough in magnitude, suggesting that they were not effective indicators of their respective latent variables. Therefore, items with standardized factor loadings smaller than .40 were removed from further analyses. One to two items were excluded from each achievement goal scale in each age group, except for the youngest age group, for

Table 2
Zero-Order Correlation Coefficients Among Observed Variables for Lower and Middle Elementary School Samples

Variable	1	2	3	4	5	6	7	8	9
1. Grade level	1.00	.02	.16	.01	−.15	−.05	.06	−.07	.15
2. Gender	−.03	1.00	−.04	.04	−.02	.09	−.10	−.06	.05
3. Mastery-approach goal	−.18	.12	1.00	.47	.14	.16	.54	−.20	.18
4. Performance-approach goal	−.35	.07	.40	1.00	.36	.34	.43	−.04	.12
5. Performance-avoidance goal	−.45	.01	.30	.58	1.00	.47	.03	.24	−.14
6. Mastery-avoidance goal	−.42	.01	.19	.45	.69	1.00	−.06	.25	−.21
7. Math self-efficacy	−.18	.12	.46	.54	.25	.17	1.00	−.27	.33
8. Help-seeking avoidance	−.09	−.21	−.10	.08	.26	.23	−.15	1.00	−.18
9. Math performance	−.12	.03	.08	.11	.04	.07	.23	−.18	1.00

Note. Coefficients for the lower elementary school sample below diagonal; coefficients for the middle elementary school sample above diagonal. Gender was coded as 1 = *boy* and 2 = *girl*. Math performance scores were teacher ratings on a scale of 1 (*below average*), 2 (*average*), and 3 (*above average*). Coefficients greater than .14 in absolute value are statistically significant at $p < .05$.

Table 3
Zero-Order Correlation Coefficients Among Observed Variables for Upper Elementary and Middle School Samples

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Grade level	1.00	.04	−.21	−.08	−.06	−.01	−.18	.10	−.01	−.15	−.07	−.02
2. Gender	−.01	1.00	−.22	−.05	.01	−.08	−.16	.05	.09	.01	−.11	.01
3. Mastery-approach goal	−.04	−.02	1.00	.42	.01	.17	.67	−.34	−.15	.44	.51	.37
4. Performance-approach goal	.08	.07	.40	1.00	.41	.40	.40	−.08	.16	.32	.26	.24
5. Performance-avoidance goal	.10	−.02	.02	.34	1.00	.45	−.09	.13	.35	.04	.04	−.16
6. Mastery-avoidance goal	.21	−.02	.07	.22	.43	1.00	−.02	.02	.47	.23	.13	−.06
7. Math self-efficacy	−.09	−.08	.67	.50	−.08	−.16	1.00	−.39	−.32	.37	.49	.56
8. Help-seeking avoidance	.01	.01	−.32	−.01	.35	.22	−.33	1.00	.20	−.31	−.46	−.29
9. Anxiety	.25	.01	−.20	.01	.38	.57	−.45	.32	1.00	.07	−.11	−.32
10. Cognitive strategy use	—	—	—	—	—	—	—	—	—	1.00	.59	.25
11. Self-regulatory strategy use	—	—	—	—	—	—	—	—	—	—	1.00	.30
12. Math performance	−.09	.03	.32	.23	−.18	−.10	.45	−.23	−.35	—	—	1.00

Note. Coefficients for the upper elementary school sample below diagonal; coefficients for the middle school sample above diagonal. Gender was coded as 1 = *boy* and 2 = *girl*. Math performance scores for the upper elementary school sample were teacher ratings on a scale of 1 (*below average*), 2 (*average*), and 3 (*above average*). Math performance scores for the middle school sample were math final exam scores on a scale of 0 to 100. Cognitive and self-regulatory strategy use variables were assessed only in the middle school sample. Coefficients greater than .16 and .09 in absolute value are statistically significant at $p < .05$ for the upper elementary and middle school samples, respectively.

which three items had to be removed from the mastery-avoidance scale. The Appendix presents factor loadings from the final four-factor model for each age group.

As Table 4 shows, the revised four-factor model based on the 2 × 2 framework (i.e., Model A) turned out to be a satisfactory approximation of the empirical data in all but the youngest age group, for which the model fit was marginal at best (NNFI = .85, CFI = .88, res. = .05). Modification indexes suggested correlating some of the error terms in each age group, which would significantly improve the model fit. When a maximum of five correlated error paths were added across the age groups, Model A was able to generate satisfactory goodness-of-fit indexes in all samples. Although the number of the correlated uniquenesses was not large, and most of them were justifiable from either a conceptual (e.g., identical wording) or a practical standpoint (e.g., adjacent items on the survey), I do not discuss the results from these correlated uniqueness models here, to avoid any possibility of making an erroneous judgment regarding the best-fitting model because of sample-specific characteristics. Therefore, on the basis of the results from Model A with no correlated uniqueness, the four-factor model was deemed an adequate representation of the achievement goal pattern for all but the youngest age group.

Goal trichotomy was tested next with different combinations of achievement goal factors. Table 4 presents fit indexes of these three-factor models. The first two trichotomous models emphasized the goal definition dimension, specifying either a single mastery or a single performance goal factor without distinguishing its approach and avoidance components. Model TA hypothesized a single mastery goal factor with separate performance-approach and performance-avoidance goals. It did not fit the data well in any of the age groups. Model TB hypothesized a single performance goal factor with independent mastery-approach and mastery-avoidance goals. This model demonstrated reasonable fit with the middle elementary school data (NNFI = .89, CFI = .91, res. = .05).

The remaining two three-factor models emphasized the goal valence dimension. Model TC specified a single approach goal factor by merging mastery-approach and performance-approach goals, whereas Model TD hypothesized a single avoidance goal factor by merging mastery-avoidance and performance-avoidance goals. Model TC with a single approach goal and separate mastery-avoidance and performance-avoidance goals again displayed marginally satisfactory fit for the middle elementary school grades (NNFI = .89, CFI = .91, res. = .05). Model TD, specifying a single avoidance goal factor with mastery-approach and performance-approach goals separated, did not fit the data well in any of the age groups. Therefore, none of the models based on goal trichotomy was able to illustrate the data pattern as well as those based on the 2 × 2 framework.

Still, the results from the series of trichotomous models associated with the youngest sample deserve some further attention. With an exception of Model TA that displayed substantially poorer fit (NNFI = .77, CFI = .80, res. = .06), compared with that of Model A (NNFI = .85, CFI = .88, res. = .05), the rest of the three-factor models demonstrated fit indexes that did not deteriorate much from those of Model A (NNFI = .85, CFI = .87, res. = .05 for Model TB; NNFI = .83, CFI = .85, res. = .06 for Model TC; NNFI = .84, CFI = .87, res. = .05 for Model TD). The set of results indicate that the children in this youngest group did not make sufficient distinction between the two performance, the two approach, or the two avoidance goals, respectively. In particular, Model TC, which merged mastery-approach and performance-approach goals into a single "approach" factor, exhibited fit indexes that were only slightly below the acceptable cutoff values in all elementary school samples. This is intriguing because the fit of the same model worsened considerably in the middle school sample.

Models based on goal dichotomy by the goal definition (i.e., Model DA) or the goal valence dimension only (i.e., Model DB) did not fit the response pattern of any age group that participated in this study. Finally, goal unidimensionality was tested by specifying a single goal factor, presumed to create variations in students' responses to all achievement goal items. This model, too, failed to depict the empirical data to a sufficient degree across all age groups.

Table 5 presents correlation coefficients among the four latent goal variables identified in Model A for each age group. Those reported by Elliot and McGregor (2001, Study 2) from their college sample are also included in the table for comparison. Although this four-factor model based on the 2 × 2 achievement goal framework was not able to reproduce the responses of the

Table 4
Goodness-of-Fit Indexes of Confirmatory Factor-Analysis Models

Model	Age group	*N*	χ^2	*df*	NNFI	CFI	Res.
2 × 2 goal framework							
A: Mastery-approach, mastery-avoidance, performance-approach, performance-avoidance	Lower elementary	214	182.91	98	.85	.88	.05
	Middle elementary	206	157.72	113	.93	.94	.04
	Upper elementary	227	243.39	164	.92	.93	.05
	Middle school	500	430.74	164	.89	.91	.05
Goal trichotomy							
TA: Mastery (approach + avoidance), performance-approach, performance-avoidance	Lower elementary	214	237.55	101	.77	.80	.06
	Middle elementary	206	242.66	116	.81	.84	.05
	Upper elementary	227	530.15	167	.62	.67	.08
	Middle school	500	1,054.27	167	.64	.68	.09
TB: Performance (approach + avoidance), mastery-approach, mastery-avoidance	Lower elementary	214	189.47	101	.85	.87	.05
	Middle elementary	206	187.85	116	.89	.91	.05
	Upper elementary	227	343.64	167	.82	.84	.06
	Middle school	500	640.27	167	.81	.83	.07
TC: Approach (mastery + performance), mastery-avoidance, performance-avoidance goals	Lower elementary	214	203.95	101	.83	.85	.06
	Middle elementary	206	190.50	116	.89	.91	.05
	Upper elementary	227	299.19	167	.86	.88	.06
	Middle school	500	679.08	167	.79	.82	.07
TD: Avoidance (mastery + performance), mastery-approach, performance-approach	Lower elementary	214	194.07	101	.84	.87	.05
	Middle elementary	206	289.21	116	.74	.78	.06
	Upper elementary	227	377.38	167	.78	.81	.06
	Middle school	500	831.89	167	.73	.76	.07
Goal dichotomy							
DA: By definition (mastery vs. performance)	Lower elementary	214	241.63	103	.77	.80	.06
	Middle elementary	206	261.83	118	.72	.79	.06
	Upper elementary	227	599.01	169	.56	.61	.08
	Middle school	500	1,267.95	169	.56	.61	.10
DB: By valence (approach vs. avoidance)	Lower elementary	214	212.18	103	.82	.84	.06
	Middle elementary	206	321.78	118	.70	.74	.07
	Upper elementary	227	422.55	169	.74	.77	.06
	Middle school	500	1,079.16	169	.63	.67	.09
Goal unidimensionality	Lower elementary	214	241.77	104	.77	.80	.06
	Middle elementary	206	374.59	119	.63	.68	.07
	Upper elementary	227	695.30	170	.47	.52	.09
	Middle school	500	1,641.90	170	.41	.47	.09

Note. NNFI = Bentler-Bonett nonnormed fit index; CFI = comparative fit index; Res. = average absolute standardized residual.

lower elementary school grade sample to an adequate degree, it was nonetheless regarded as the best empirical representation among the models tested for this youngest group, because satisfactory model fit was achievable with correlated uniquenesses.

The correlation coefficients between any two achievement goal factors in the present study tended to decrease in magnitude as the age of the samples increased. None of the achievement goal correlation in the middle elementary, upper elementary, and middle school samples appeared so high as to pose a threat to discriminant validity of the goal factors (ϕs $\leq$ .67). In contrast, several correlation coefficients at the lower elementary school grades were quite large. Specifically, performance-approach and performance-avoidance (ϕ = .88), mastery-

Table 5
Correlation Coefficients Among Latent Goal Variables

	Present study				Elliot and McGregor, (2001), study 2
Achievement goal	Lower elementary: Grades 1 and 2	Middle elementary: Grades 3 and 4	Upper elementary: Grades 5 and 6	Middle school: Grades 7–9	College
Mapp—Papp	.81	.59	.67	.45	−.14
Mapp—Pavd	.63	.26	.14 *ns*	−.02 *ns*	−.08
Mapp—Mavd	.40	.20 *ns*	.25	.21	.37
Papp—Pavd	.88	.67	.48	.53	.18
Papp—Mavd	.83	.55	.35	.48	.04
Pavd—Mavd	.87	.52	.42	.45	.27

Note. Mapp = mastery-approach; Papp = performance-approach; Pavd = performance-avoidance; Mavd = mastery-avoidance; *ns* = statistically nonsignificant at $p < .05$. Correlation coefficients are from Model A.

approach and performance-approach (ϕ = .81), and mastery-avoidance and performance-avoidance goals (ϕ = .87) of this youngest elementary school sample were highly correlated, indicating the children in Grades 1 and 2 in the present study did not differentiate these achievement goals as clearly as did their older counterparts. These results are consistent with an earlier observation that the overall fits of the three trichotomous models, Models TB, TC, and TD, respectively, were generally acceptable in the youngest age group.

On average, the correlation coefficients obtained in the present study with elementary and middle school students were larger than those reported by Elliot and McGregor (2001) with college students, except for those between mastery-approach and mastery-avoidance goals. The performance-approach and the mastery-avoidance goal factors also demonstrated sizeable correlation across the age groups, especially at the lower elementary school grades (ϕ = .83).

Testing Age-Related Differences in the Strengths of Achievement Goal Endorsement

A MANOVA was run on the set of four achievement goals with age group and gender as between-subjects factors. I used Pillai's trace instead of Wilks's lambda to determine multivariate significance, because the assumption of homogeneity of covariance matrices was not met according to the Box's *M* test (Stevens, 1992; Tabachnick & Fidell, 2001). Statistically significant differences existed by age group, $F(12, 3411) = 45.30$, $p < .001$ (partial η^2 = .14) but not gender. The interaction between age group and gender was significant, $F(12, 3411) = 2.15$, $p < .05$ (partial η^2 = .01). Table 6 reports the mean achievement goal scores used in the MANOVA and subsequent paired-samples *t* tests. These values were computed with a smaller number of items, as determined in the preceding CFA (see Appendix) and, hence, differ somewhat from those reported in Table 1. However, the results were almost identical, regardless of whether all 23 items or only those with substantial factor loadings were included in the analyses.

Tests of between-subjects effects detected statistically significant age-related differences on all four achievement goals, $F(3, 1138) = 106.71$, $p < .001$ (partial η^2 = .22) for mastery-approach; $F(3, 1138) = 17.84$, $p < .001$ (partial η^2 = .05) for performance-approach; $F(3, 1138) = 99.05$, $p < .001$ (partial η^2 = .21) for performance-avoidance; and $F(3, 1138) = 12.96$, $p < .001$ (partial η^2 = .03) for mastery-avoidance goals. The post hoc Scheffé procedures at $p < .001$ revealed that the mean mastery-approach goal score of the lower elementary school grades (M = 4.36) was statistically higher than that of the middle elementary school grades (M = 3.95), which, in turn, was statistically higher than those of the upper elementary (M = 3.46) and the middle school samples (M = 3.22). The mastery-approach goal scores of the latter two older age groups were not statistically different. The same pattern held with regard to the performance-avoidance goal. The mean performance-avoidance goal score was the highest in the lower elementary school grades (M = 3.59), next highest in the middle elementary school grades (M = 2.92), and lowest in the upper elementary (M = 2.43) and the middle school samples (M = 2.27). The scores of the latter two groups did not differ from each other.

In terms of the performance-approach goal, only the mean score of the youngest group was statistically different from that of the oldest group in the present study. The performance-approach goal score of the students in Grades 1 and 2 (M = 3.92) was statistically higher than that of the middle school students (M = 3.41). The mean mastery-avoidance goal score was the highest among the students in the middle elementary school grades (M = 3.32), which was statistically higher than those of the upper elementary (M = 2.81) and the middle school samples (M = 2.88). It is noteworthy that there was no statistically significant difference in any of the mean achievement goal scores between the upper elementary and the middle school samples.

The Age Group × Gender interaction was significant only on the mastery-approach goal. The mean mastery-approach goal score of the boys decreased as the age of the samples increased but leveled out between the upper elementary (M = 3.48) and the middle school years (M = 3.42). In contrast, the mean mastery-approach goal score of the girls continued to decrease as the age of the samples increased from the lower elementary school grades to the middle school years.

It is difficult to claim that the differences between the age groups reported so far entirely represent actual variations in achievement goal endorsement, because the younger groups tended to provide higher achievement goal ratings than did the older groups, regardless of the goal valence or the goal definition. The group differences likely owe to not only genuine changes in the strength of each achievement goal by age but also general developmental characteristics in survey responding. To probe whether the students at different grade levels endorsed a particular achievement goal more strongly than they did others, I performed multiple paired-samples *t* tests at $p < .001$ for all possible pairs of achievement goals within each age group.

All paired comparisons proved statistically significant. Table 6 presents the summary. The students in the two younger age groups

Table 6
Summary of MANOVA and Paired-Samples t Tests

	Mean achievement goal score								
Age group	Mapp	*SD*	Papp	*SD*	Pavd	*SD*	Mavd	*SD*	Summary of paired-samples *t* tests
Lower elementary	4.36_{a}	.65	3.92_{a}	1.08	3.59_{a}	1.17	$3.17_{a,b}$	1.39	Mapp > Papp > Pavd > Mavd
Middle elementary	3.95_{b}	.84	$3.71_{a,b}$	.95	2.92_{b}	1.20	3.32_{a}	1.14	Mapp > Papp > Mavd > Pavd
Upper elementary	3.46_{c}	.82	$3.69_{a,b}$	.80	2.43_{c}	.87	2.81_{b}	.98	Papp > Mapp > Mavd > Pavd
Middle school	3.22_{c}	.92	3.41_{b}	.85	2.27_{c}	.84	2.88_{b}	.88	Papp > Mapp > Mavd > Pavd

Note. Mapp = mastery-approach; Papp = performance-approach; Pavd = performance-avoidance; Mavd = mastery-avoidance. Different subscripts denote a statistically significant difference within each column at $p < .001$.

(i.e., Grades 1–4) expressed the strongest agreement with the mastery-approach goal items, followed by the performance-approach goal items. The two avoidance goals received either the lowest or the second lowest ratings in these two younger samples. In comparison, the students in the two older age groups, the upper elementary and the middle school samples (i.e., Grades 5–9), provided the highest endorsement ratings on the performance-approach goal items, followed by the mastery-approach, the mastery-avoidance, and the performance-avoidance goal items in the same descending order.

Testing Relationships of Achievement Goals With Other Constructs

A full CFA model with all latent variables, including the four achievement goal factors identified from Model A in previous CFAs, was fitted to the data within each age group. CFA was deemed a more appropriate strategy than structural equation modeling for testing achievement goal relationships, because either insufficient or contradictory theoretical accounts exist at present regarding temporal and causal predominance among the variables. In these CFA models, uniquenesses of several indicators were allowed to covary, as the objective was no longer testing the validity of a particular structure, and obtaining unbiased construct relations was deemed more important. To minimize chance findings, only those theoretically or empirically justifiable and projecting significant drops in chi-square values at $p < .001$ were incorporated. A vast majority of the correlated uniquenesses were between items that contained the same phrase or successively appeared on the survey. The cognitive and self-regulatory strategy use variables in the middle school sample were respecified as a single latent variable, because the students did not distinguish between these two strategy use variables ($\phi = .96$).

The final CFA models were able to reproduce the data to satisfactory degrees in all age groups, $\chi^2(187, N = 214) = 269.78$, $p < .001$ (NNFI = .90, CFI = .92, res. = .05) for the lower elementary; $\chi^2(325, N = 206) = 465.98$, $p < .001$ (NNFI = .89, CFI = .90, res. = .05) for the middle elementary; $\chi^2(497, N = 227) = 635.49$, $p < .001$ (NNFI = .93, CFI = .94, res. = .05) for the upper elementary; and $\chi^2(940, N = 500) = 1{,}554.84$, $p < .001$ (NNFI = .90, CFI = .91, res. = .04) for the middle school sample. Table 7 presents correlation coefficients among the latent variables.

A mastery-approach goal exhibited strong positive correlation with self-efficacy, which proved to be statistically significant in all age groups, with the coefficients ranging between .58 and .86. It demonstrated negative correlation with help-seeking avoidance with the coefficient being statistically significant only in the middle school sample ($\phi = -.22$). There was no significant correlation between a mastery-approach goal and anxiety. A mastery-approach goal also significantly and positively correlated with strategy use in the middle school sample ($\phi = .65$), as well as math performance in the upper elementary ($\phi = .28$) and the middle school samples ($\phi = .40$).

A performance-approach goal displayed a correlation coefficient of unity with self-efficacy in the lower elementary school sample, which indicates that the children in this youngest age group did not distinguish their responses toward these two variables. In fact, the responses of these young children correlated very strongly across all variables assessed in the present study, including achievement goals. Math self-efficacy hence demonstrated strong positive correlation with all achievement goals for this youngest group, with the coefficients being noticeably larger when the achievement goals were of positive (ϕs = .73 and 1.00 with a mastery-approach and a performance-approach goal, respectively), rather than negative valence (ϕs = .55 and .42 with a performance-avoidance and a mastery-avoidance goal, respectively).

Whereas a performance-approach goal did not correlate significantly with help-seeking avoidance, it correlated significantly and

Table 7
Correlation Coefficients of Latent Goal Variables With Other Variables

Variable	Mastery-approach	Performance-approach	Performance-avoidance	Mastery-avoidance
Math self-efficacy				
Lower elementary	.73	1.00	.55	.42
Middle elementary	.58	.65	.08 *ns*	.08 *ns*
Upper elementary	.86	.68	.07 *ns*	.02 *ns*
Middle school	.74	.42	−.11 *ns*	.02 *ns*
Help-seeking avoidance				
Middle elementary	−.20 *ns*	−.09 *ns*	.19	.18 *ns*
Upper elementary	−.10 *ns*	.09 *ns*	.47	.23
Middle school	−.22	.06 *ns*	.16	.05 *ns*
Anxiety				
Upper elementary	.01 *ns*	.24	.64	.84
Middle school	−.12 *ns*	.35	.39	.67
Strategy use				
Middle school	.65	.48	.08 *ns*	.31
Math performance[a]				
Lower elementary	.08 *ns*	.19	.03 *ns*	−.03 *ns*
Middle elementary	.16 *ns*	−.03 *ns*	−.12 *ns*	−.23
Upper elementary	.28	.13 *ns*	−.17	−.04 *ns*
Middle school	.40	.19	−.18	−.10 *ns*

Note. *ns* = statistically non-significant at $p < .05$. Correlation coefficients are from the 2 × 2 achievement goal CFA models to which all latent variables were added. Correlated uniquenesses were allowed.
[a] Teacher ratings for elementary school students; final exam scores for middle school students.

positively with anxiety (ϕs = .24 and .35 in the upper elementary and the middle school samples, respectively). It also correlated positively with strategy use (ϕ = .48), although this coefficient was smaller in magnitude compared with the one between a mastery goal and strategy use. Math performance scores correlated positively with a performance-approach goal in the lower elementary and the middle school samples (both ϕs = .19).

With an exception of its positive correlation with self-efficacy in the youngest age group, a performance-avoidance goal demonstrated nonsignificant to negative associations with adaptive variables and strong positive associations with maladaptive variables. Specifically, it did not correlate significantly with self-efficacy, again except for the students in the lower elementary school grades, or strategy use but correlated negatively with math performance in the upper elementary (ϕ = −.17) and the middle school samples (ϕ = −.18). Instead, a performance-avoidance goal correlated positively with help-seeking avoidance (.16 ≤ ϕs ≤ .47) in the middle elementary, the upper elementary, and the middle school samples and anxiety (ϕs = .64 and .39) in the two oldest samples in the present research.

A mastery-avoidance goal displayed relationships that were, on the whole, most similar to those associated with a performance-avoidance goal, yet different from those on some important aspects. Its mostly nonsignificant correlations with self-efficacy and strong positive correlation with anxiety (ϕs = .84 and .67) are analogous to the pattern associated with a performance-avoidance goal. However, the positive correlations with help-seeking avoidance and negative correlation with math performance in the two older samples of a mastery-avoidance goal were weaker in magnitude, compared with those of a performance-avoidance goal. A mastery-avoidance goal's relationships with these variables were thus largely nonsignificant, except in the upper elementary school sample, with regard to help-seeking avoidance (ϕ = .23), and the middle elementary school sample, with regard to math performance (ϕ = −.23). It is interesting to note that a mastery-avoidance goal demonstrated significant positive correlation with strategy use (ϕ = .31), as did mastery-approach and performance-approach goals, although this correlation was the weakest in size among the three.

Finally, among all the variables assessed in the present study, math self-efficacy was the only variable that displayed an unwavering significant relationship with math performance in all age groups. The correlation between self-efficacy and performance tended to become stronger as the students became older, with the coefficients of .21, .35, .34, and .51 in the lower elementary, the middle elementary, the upper elementary, and the middle school samples, respectively.

Discussion

Validity of the 2 × 2 Achievement Goal Framework for School-Aged Children

The present results provided empirical support for the 2 × 2 achievement goal framework proposed by Elliot (1999) and Pintrich (2000), at least for the Korean children and adolescents in Grades 3–9 that participated in this research. When models hypothesizing a different set of achievement goals were pitted against each other, the one that differentiated achievement goals by both the valence of achievement striving and the definition of competence best described these students' response pattern to the achievement goal survey. This study is one of the few that involved school-aged children and adolescents, unlike most existing studies on the 2 × 2 framework that relied upon college samples. Therefore, it is an important finding that the children as young as third graders in elementary schools did make some distinctions between the different purposes they might try to fulfill when engaging in achievement-related behaviors in math and the different criteria they might use for judging their math competence, when asked to do so.

Several developmental trends are worthy of note. Younger children did not make as clear differentiation of the multiple achievement goals as did their older counterparts. Whereas only the four-goal model adequately described the achievement goal responses of the Korean middle school students (i.e., roughly equivalent to U.S. Grades 7–9), some of the models with fewer numbers of achievement goals were able to illustrate the achievement goal patterns of the younger children to reasonable degrees. In particular, the three-factor model specifying a single approach goal without splitting it into mastery-approach and performance-approach components displayed fit indexes almost comparable to those of the four-factor model in the elementary school samples. In fact, this model was able to reproduce the elementary school children's achievement goal responses to satisfactory levels when correlated uniquenesses were allowed.

These results are consistent with the supposition that young children would more clearly distinguish between situations of potential successes and those of potential failures than they would success situations defined by different types of competence. Young children, who have not yet fully grasped the meaning of normative competence (Nicholls, 1984; Ruble et al., 1980; Stipek & Mac Iver, 1989), would find it more challenging to tell apart successes achieved by task mastery and those defined by relative superiority.

However, one finding is in seeming conflict with these otherwise plausible conjectures. The trichotomous model proposing a single performance goal without bifurcating it into its approach and avoidance components also exhibited decent fit to the achievement goal responses of the elementary school children. This model, too, was able to illustrate the data satisfactorily when correlated uniquenesses were introduced. Although the positive correlation between performance-approach and performance-avoidance goals has been a consistent finding in the literature (Bong, 2001; Elliot & Church, 1997; Middleton & Midgley, 1997; Skaalvik, 1997), the correlation was never as strong as to be near identity. Considering that middle school students had been the youngest participants in previous studies, and the single-performance-goal model demonstrated good fit only in the elementary school samples in the present study, it seems safe to conclude that this finding has to do with the age of the respondents.

Elliot, Conroy, and their colleagues suggested that individual temperament or personality disposition plays an important role in guiding individuals to certain types of achievement goals. Fear of failure repeatedly emerged as a common antecedent of both performance-approach and performance-avoidance goals in their research (Conroy & Elliot, 2004; Conroy, Elliot, & Hofer, 2003; Elliot & Church, 1997; Elliot & McGregor, 2001). In another study, approach temperament consisted of extraversion, positive emotionality, and behavioral activation predicted mastery-approach and performance-approach goals, whereas avoidance temperament consisted of neuroticism, negative emotionality, and behavioral inhibition predicted performance-approach and performance-avoidance goals (Elliot & Thrash, 2002). Because a per-

formance-approach goal is jointly determined by both approach and avoidance temperaments, it is not surprising that a single "performance" factor was able to encapsulate both the approach and the avoidance tendencies to a certain extent.

Still, this does not explain why the performance-approach and the performance-avoidance goals should display particularly strong correlation among younger children. Two speculations are offered. First, older students are presumed to have stored sufficient success and failure experiences in their self-schema. This achievement history is believed to guide them to either approach or avoidance paths in given achievement situations. Young children, with only limited amount of such information at their disposal, might be swayed more strongly by personality factors when it comes to adopting an achievement goal. The two performance goals originating from similar personality disposition thus correlate more strongly with each other.

Second, younger children's motivation is more heavily affected by desire to please significant adults, such as parents, than is older children's motivation (Mac Iver, Stipek, & Daniels, 1991). Asian students, Korean students included, are further known for their keen interest in others' evaluations and reactions toward them, as well as their willingness to conform to the norm in their social network (Markus & Kitayama, 1991; Oishi & Diener, 2001). Both performance goals, as assessed in this study, are largely made up of two components: validation of competence and normative comparison of ability. The eagerness of young Korean children to satisfy their parents and teachers and have their ability validated would render the two performance goals to covary more strongly with each other among this population. Whether performance goals operationalized in different ways (e.g., outcome goals, see Grant & Dweck, 2003) should also exhibit this developmental pattern remains to be seen.

Age-Related Variations in the Strengths of Achievement Goals

A mastery-approach goal was most strongly endorsed by the youngest group of children in this study as the primary reason for doing math work. However, these young children also expressed stronger endorsements to all other achievement goals, compared with their older counterparts. On one hand, it seems most reasonable to view these between-group differences as a consequence of young children's tendency to provide higher ratings on survey items, rather than authentic differences in the strengths of each achievement goal across the age groups. On the other hand, within-group comparisons of the achievement goal strengths should be able to shed light on the developmental trends in achievement goal adoption, as everyone within each age group would have been exposed to similar age-specific extraneous factors.

The results from the within-group comparisons supported the hypotheses generated on the basis of developmental research. Specifically, the younger children in Grades 1–4 in the present study indicated that they pursued a mastery-approach goal most strongly, followed by a performance-approach goal. The two avoidance goals received significantly lower average ratings than did the two approach goals. Because young children subscribe to an incremental theory of intelligence (Dweck & Leggett, 1988), evaluate their own competence in more absolute than normative terms (Harter, 1975, 1998; Ruble et al., 1980), and tend to function in learning environments that emphasize task mastery over comparative superiority (Eccles et al., 1993), a mastery-approach goal would best represent the reasons why they engage in achievement behaviors in school. It also makes sense that a performance-approach goal received the next highest ratings from these young children, given their tendency to appraise most achievement situations as appetitive rather than aversive (Stipek & Mac Iver, 1989).

In contrast to the younger children, the two older groups of students in Grades 5–9 rated the performance-approach goal the highest. This shift is noteworthy because it emerged despite the typical tendency of respondents to express the strongest agreement to mastery-approach goal items out of self-enhancement and impression-management concerns (Day, Radosevich, & Chasteen, 2003). Respondents are also not willing to admit the degree to which they rely on social comparison for evaluating their own competence because of their apprehensions about social desirability (Harter et al., 1992). The strongest endorsement ratings for the performance-approach goal by the students in the upper elementary and the middle school grades thus indicate a quite substantial difference in the degree to which these students preferred a performance-approach goal to the other achievement goals.

Upper elementary school grades in the present study refer to Grades 5 and 6. Fifth graders in Korean elementary school are 10 to 11 years of age, depending on their birth month. Researchers proposed that it is around these ages that children begin to acquire differentiated conceptions of ability (Nicholls, 1984; Stipek & Mac Iver, 1989) and use social comparative information for evaluating the quality of their work (Butler, 1989; Ruble et al., 1980). It is also during these last 2 years of elementary school that competition with peers and relative ability concerns gradually rise to the surface in Korean schools. When students enter middle school, they encounter a considerably more heavily ability-focused learning environment, the same way U.S. students do (Eccles et al., 1993; Stipek & Mac Iver, 1989). It may be of no surprise, therefore, that the stronger and more ubiquitous focus on ability and grades in the upper elementary and the middle school environments materialized as stronger personal performance-approach goals for these students (Ames & Archer, 1988; Harter et al., 1992; Midgley et al., 1995).

Age-Related Variations in the Associations of Achievement Goals

With few exceptions, the results are generally consistent with the literature and across the current age groups on the adaptive nature of a mastery-approach goal and the maladaptive nature of a performance-avoidance goal. A mastery-approach goal correlated positively with self-efficacy, strategy use, and performance in math and negatively with help-seeking avoidance and anxiety. A performance-avoidance goal showed almost a mirror pattern of relationships to those of a mastery-approach goal. A performance-approach goal showed mixed relationships with positive and negative motivational variables, with significant relationship with math performance in only the middle school sample.

So far, age-related differences in the differentiation and the strengths of achievement goals indicate that clear differences exist between the two younger and the two older groups of students participating in this research. The finding that a performance-approach goal starts predicting students' math performance from the middle school years is especially intriguing. It indicates that the adaptive nature of a performance-approach goal, advocated by the

revised goal theorists (Harackiewicz, Barron, Pintrich, Elliot, & Thrash, 2002) and demonstrated repeatedly among college students (e.g., Harackiewicz, Barron, Tauer, Carter, & Elliot, 2000; Harackiewicz, Barron, Tauer, & Elliot, 2002), might begin to set in motion only after children learn to recognize and appreciate the potential benefits of achieving success with less effort and only in environments where outperforming peers gets rewarded in the form of better grades.

More interesting, it is also during the middle school years that the relatively more positive quality of a mastery-approach goal, compared with that of a performance-approach goal, became more evident. Urdan and Midgley (2003) argued that what may be of greater consequence for students' motivation in school is the positive impact of a mastery focus in the learning environment than the negative impact of an ability focus. The current results seem to support their claim and extend it by suggesting that personal mastery-approach goals may be particularly conducive in cutting down potential maladaptive tendencies, such as help-seeking avoidance and test anxiety, among early adolescents in a performance-oriented learning environment. Analogous results were reported in Conroy and Elliot (2004) with undergraduate students in a large university. The participants responded to fear of failure and achievement goal surveys four times throughout physical activity courses, with successive testing conducted after 2 days (T_2), 1 week (T_3), and 3 weeks (T_4), following an initial testing (T_1). Contrary to the researchers' theoretical predictions, a mastery-approach goal assessed at T_3 appeared as a negative antecedent of fear of failure at T_4. As the investigators noted, this relationship was significant only between the third and the fourth measurements and was quite negligible in magnitude. Such limitations notwithstanding, the researchers acknowledged that a mastery-approach goal might serve a protective and "developmentally adaptive" function of inhibiting unpleasant self-conscious emotions such as shame, which eventuate in fear of failure.

The present results with much younger children and adolescents correspondingly suggest that, although the relationships of their mastery-approach and performance-approach goals with positive psychological and behavioral outcomes may look similar at first glance, it is the mastery-approach goal that provides a stronger protective shield or "psychological armor" that helps them ward off harmful thoughts and affects. Moreover, this phenomenon appears to play out more vividly among students who are confronted with heavily competitive and ability-focused learning environments (see also Turner et al., 2002).

In sum, the strong relationships consistently demonstrated by a performance-approach goal with a performance-avoidance goal, as well as a host of unhealthy variables such as anxiety (Bong, 2005; Bong & Kim, 2006; Conroy et al., 2003; Elliot & Church, 1997; Elliot & McGregor, 2001; Middleton & Midgley, 1997; Midgley & Urdan, 2001; Ross et al., 2002), once again highlight the critical importance of providing mastery-oriented learning environments to school-aged children and adolescents. This conclusion appears highly justifiable, given the tight relationships between school and classroom goal structures and students' personal achievement goals (Church et al., 2001; Midgley et al., 1995; Roeser et al., 1996; Turner et al., 2002; Wolters, 2004) and the particularly devastating impact of ability-oriented learning environments on students with low perceived competence (Dweck, 1986; Elliott & Dweck, 1988; Jagacinski et al., 2001).

Case for the Mastery-Avoidance Achievement Goal

The present research was set out to test, among other things, whether a mastery-avoidance goal should be considered as a part of legitimate representations of school-aged children's underlying motives in achievement situations. As discussed earlier, the results from confirmatory factor analyses supported empirical independence of a mastery-avoidance goal, as the four-goal model best described the achievement goal responses of all age groups. The mastery-avoidance goal factor so identified, however, did not covary much with variables of either adaptive or maladaptive nature. On the whole, the results from the present investigation seem to support Elliot and McGregor's (2001) contention regarding the motivational quality of a mastery-avoidance goal. Whereas students with stronger performance-avoidance goals clearly felt more anxious, demonstrated stronger tendencies to avoid seeking necessary help, and performed more poorly in math, those with stronger mastery-avoidance goals did not always follow this maladaptive pattern. In particular, as students reported stronger mastery-avoidance goals, they also reported greater use of cognitive and self-regulatory strategies in math.

Nevertheless, there still remain unresolved questions on the exact makeup of a mastery-avoidance goal. The strong positive correlation between mastery-avoidance and performance-approach goals observed in previous studies (Conroy et al., 2003; Elliot & McGregor, 2001, Study 1; Finney et al., 2004) turned up again in this investigation, despite the fact they share neither the goal valence nor the goal definition. The mastery-avoidance goal also showed the strongest correlation with math anxiety, compared with the other achievement goals. Inclusion of such phrases as "I'm afraid," "I'm concerned," or "I worry" in the items might have contributed to this relationship. However, previous studies reported similar findings with college students in sport contexts. The mastery-avoidance goals of these students displayed stronger correlation with fear of failure than did their performance-avoidance goals. The differences in the magnitude of these correlation coefficients were slight yet reliable across multiple waves (Conroy & Elliot, 2004; Conroy et al., 2003). These findings suggest that a mastery-avoidance goal may be more strongly guided by fear of failure than previously assumed.

Some have raised skepticism about the need for an additional achievement goal, such as the mastery-avoidance goal, when the exact meanings and definitions of existing goals have yet to be agreed upon (see, e.g., Brophy, 2005; Grant & Dweck, 2003). The present results at least appear to justify additional research on the mastery-avoidance goal. The most pressing need appears to be elucidating what constitute the essential forbearers and outcomes of a mastery-avoidance goal. Further research on the nomological network of achievement goals should be able to clarify this.

Contributions, Limitations, and Future Directions

In this study, several achievement goal items that had been in use with U.S. students had to be dropped because they failed to demonstrate adequate loadings on their respective latent goal variables. Some of these items are judged to be not sufficiently

pertinent to regular Korean classroom situations. For example, it would have been difficult for Korean children to fully identify with the statement, "I would feel really good if I were the only one who could answer the teachers' questions in my math class," because typical Korean classroom instruction is characterized by one-way communication from the teacher to the whole class (Bong, 2003; Bong & Kim, 2006). Other reasons appear more developmental in nature. For instance, items with double negatives (e.g., "I'm afraid that I won't do my very best in my math class," "It is important to me 'not' to do my math work incorrectly") were excluded from analyses for the youngest group of children. Even so, it is unlikely that discrepancies in survey items were responsible for any of the main conclusions from the present investigation because the findings were generally consistent with the existing literature.

Pintrich (2003) posed seven important questions that need to be addressed in future motivation studies. Among those, this investigation provides some answers to the following two questions from an achievement goal perspective: "What motivates students in the classrooms?" and "How does motivation change and develop?" Unfortunately, the regulation of achievement goals, such as whether learners pursuing a certain achievement goal are more likely to turn to a different achievement goal when circumstances change, could not be tested in this study. There has been a proposal that students with a strong performance-approach goal would quickly switch to a performance-avoidance goal when they experience failure or when their learning environments become increasingly challenging (Bong, 2005; Brophy, 2005). It is also possible that, as learners become highly familiar with given tasks, their initial interest toward the tasks may start to plummet. After a certain level of mastery is achieved, they may begin to pursue a different achievement goal.

The ultimate question, in the end, is how faithfully the achievement goal items portray the psychological reality of these young children and adolescents (Brophy, 2005). Up until now, all supporting evidence for the mastery-avoidance goal has been confined to survey responses. Attempts to integrate self-report questionnaires with more naturalistic approaches, such as interviews and classroom observations (see, e.g., Turner et al., 2002), thus appear particularly promising. It will be interesting to see if young children readily cite a mastery-avoidance motive as the primary reason behind their achievement-oriented behaviors in the classroom.

References

Ames, C. (1992). Classrooms: Goals, structure, and student motivation. *Journal of Educational Psychology, 84,* 261–271.

Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology, 80,* 260–267.

Anderman, E. M., & Midgley, C. (1997). Changes in achievement goal orientations, perceived academic competence, and grades across the transition to middle-level schools. *Contemporary Educational Psychology, 22,* 269–298.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bong, M. (2001). Between- and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task-value, and achievement goals. *Journal of Educational Psychology, 93,* 23–34.

Bong, M. (2003). Choices, evaluations, and opportunities for success: Academic motivation of Korean adolescents. In F. Pajares & T. C. Urdan (Eds.), *Adolescence and education: Vol. 3. International perspectives* (pp. 323–345). Greenwich, CT: Information Age.

Bong, M. (2005). Within-grade changes in Korean girls' motivation and perceptions of the learning environment across domains and achievement levels. *Journal of Educational Psychology, 97,* 656–672.

Bong, M. (2008). Effects of parent-child relationships and classroom goal structures on motivation, help-seeking avoidance, and cheating. *Journal of Experimental Education, 76,* 191–217.

Bong, M., & Clark, R. E. (1999). Comparison between self-concept and self-efficacy in academic motivation research. *Educational Psychologist, 34,* 139–154.

Bong, M., & Kim, S. (2006). Korean students' reactions to perceived learning environment, parental expectations, and performance feedback. In D. McInerney, M. Dowson, & S. Van Etten (Eds.), *Research on sociocultural influences on motivation and learning: Vol. 6. Effective schools* (pp. 235–262). Greenwich, CT: Information Age.

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review, 15,* 1–40.

Brophy, J. (2005). Goal theorists should move on from performance goals. *Educational Psychologist, 40,* 167–176.

Butler, R. (1989). Mastery versus ability appraisal: A developmental study of children's observations of peers' work. *Child Development, 60,* 1350–1361.

Church, M. A., Elliot, A. J., & Gable, S. L. (2001). Perceptions of classroom environment, achievement goals, and achievement outcomes. *Journal of Educational Psychology, 93,* 43–54.

Conroy, D. E., & Elliot, A. J. (2004). Fear of failure and achievement goals in sport: Addressing the issue of the chicken and the egg. *Anxiety, Stress, and Coping, 17,* 271–285.

Conroy, D. E., Elliot, A. J., & Hofer, S. M. (2003). A 2 × 2 achievement goals questionnaire for sport: Evidence for factorial invariance, temporal stability, and external validity. *Journal of Sport & Exercise Psychology, 25,* 456–476.

Day, E. A., Radosevich, D. J., & Chasteen, C. S. (2003). Construct- and criterion-related validity of four commonly used goal orientation instruments. *Contemporary Educational Psychology, 28,* 434–464.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41,* 1040–1048.

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development.* Philadelphia, PA: Taylor & Francis.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95,* 256–273.

Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development, 64,* 830–847.

Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist, 34,* 169–189.

Elliot, A. J. (2005). A conceptual history of the achievement goal construct. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 52–72). New York: Guilford Press.

Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 72,* 218–232.

Elliot, A. J., & Harackiewicz, J. M. (1994). Goal setting, achievement orientation, and intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology, 66,* 968–980.

Elliot, A. J., & Harackiewicz, J. M. (1996). Approach and avoidance achievement goals and intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology, 70,* 461–475.

Elliot, A. J., & McGregor, H. A. (2001). A 2 × 2 achievement goal

framework. *Journal of Personality and Social Psychology, 80,* 501–519.

Elliot, A. J., & Thrash, T. M. (2002). Approach-avoidance motivation in personality: Approach and avoidance temperaments and goals. *Journal of Personality and Social Psychology, 82,* 804–818.

Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54,* 5–12.

Finney, S. J., Pieper, S. L., & Barron, K. E. (2004). Examining the psychometric properties of the Achievement Goal Questionnaire in a general academic context. *Educational and Psychological Measurement, 64,* 365–382.

France-Kaatrude, A. C., & Smith, W. P. (1985). Social comparison, task motivation, and the development of self-evaluative standards in children. *Developmental Psychology, 21,* 1080–1089.

Grant, H., & Dweck, C. S. (2003). Clarifying achievement goals and their impact. *Journal of Personality and Social Psychology, 85,* 541–553.

Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology, 94,* 638–645.

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M., & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology, 92,* 316–330.

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology, 94,* 562–575.

Harter, S. (1975). Developmental differences in the manifestation of mastery motivation on problem-solving tasks. *Child Development, 46,* 370–378.

Harter, S. (1990) Causes, correlates, and the functional role of global self-worth: A life-span perspective. In R. J. Sternberg & J. Kolligian (Eds.), *Competence considered* (pp. 67–97). New Haven, CT: Yale University Press.

Harter, S. (1998). The development of self-representations. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (5th ed., pp. 553–617). New York: Wiley.

Harter, S., Whitesell, N. R., & Kowalski, P. (1992). Individual differences in the effects of educational transitions on young adolescents' perceptions of competence and motivational orientation. *American Educational Research Journal, 29,* 777–807.

Hinkle, D. E., Wiersma, W. W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin.

Jagacinski, C. M., Madden, J. L., & Reider, M. H. (2001). The impact of situational and dispositional achievement goals on performance. *Human Performance, 14,* 321–337.

Joo, Y. J., Bong, M., & Choi, H. J. (2000). Self-efficacy for self-regulated learning, academic self-efficacy, and Internet self-efficacy in Web-based instruction. *Educational Technology Research and Development, 48*(2), 5–18.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford Press.

Linnenbrink, E. A. (2005). The dilemma of performance-approach goals: The use of multiple goal contexts to promote students' motivation and learning. *Journal of Educational Psychology, 97,* 197–213.

Mac Iver, D. J., Stipek, D. J., & Daniels, D. H. (1991). Explaining within-semester changes in students' effort in junior high school and senior high school courses. *Journal of Educational Psychology, 83,* 201–211.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98,* 224–253.

Marsh, H. W., Craven, R. G., & Debus, R. (1991). Self-concepts of young children 5 to 8 years of age: Measurement and multidimensional structure. *Journal of Educational Psychology, 83,* 377–392.

Middleton, M. J., & Midgley, C. (1997). Avoiding the demonstration of lack of ability: An underexplored aspect of goal theory. *Journal of Educational Psychology, 89,* 710–718.

Midgley, C., Anderman, E., & Hicks, L. (1995). Differences between elementary and middle school teachers and students: A goal theory approach. *Journal of Early Adolescence, 15,* 90–113.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., et al. (2000). *Manual for the Patterns of Adaptive Learning Scales.* Ann Arbor: University of Michigan Press.

Midgley, C., & Urdan, U. (2001). Academic self-handicapping and achievement goals: A further examination. *Contemporary Educational Psychology, 26,* 61–75.

Nicholls, J. G. (1984). Conceptions of ability and achievement motivation. In R. Ames & C. Ames (Eds.), *Research on motivation in education: Vol. 1. Student motivation* (pp. 39–73). Orlando, FL: Academic Press.

Oishi, S., & Diener, E. (2001). Goals, culture, and subjective well-being. *Personality and Social Psychology Bulletin, 27,* 1674–1682.

Pastor, D. A., Barron, K. E., Miller, B. J., & Davis, S. L. (2004, April). *College students' achievement goal orientation profiles.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Pintrich, P. R. (2000). An achievement goal theory perspective on issues in motivation terminology, theory, and research. *Contemporary Educational Psychology, 25,* 92–104.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95,* 667–686.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82,* 33–40.

Roeser, R. W., Midgley, C., & Urdan, T. C. (1996). Perceptions of the school psychological environment and early adolescents' psychological and behavioral functioning in school: The mediating role of goals and belonging. *Journal of Educational Psychology, 88,* 408–422.

Ross, M. E., Shannon, D. M., Salisbury-Glennon, J. D., & Guarino, A. (2002). The Patterns of Adaptive Learning Survey: A comparison across grade levels. *Educational and Psychological Measurement, 62,* 483–497.

Ruble, D. N., Boggiano, A. K., Feldman, N. S., & Loebl, J. H. (1980). Developmental analysis of the role of social comparison in self-evaluation. *Developmental Psychology, 16,* 105–115.

Ruble, D. N., Feldman, N. S., & Boggiano, A. K. (1976). Social comparison between young children in achievement situations. *Developmental Psychology, 12,* 192–197.

Ryan, A. M., Gheen, M. H., & Midgley, C. (1998). Why do some students avoid asking for help? An examination of the interplay among students' academic self-efficacy, teachers' social-emotional role, and the classroom goal structure. *Journal of Educational Psychology, 90,* 528–535.

Ryan, A. M., & Pintrich, P. R. (1997). "Should I ask for help?" The role of motivation and attitudes in adolescents' help seeking in math class. *Journal of Educational Psychology, 89,* 329–341.

Senko, C., & Harackiewicz, J. M. (2002). Performance goals: The moderating roles of context and achievement orientation. *Journal of Experimental Social Psychology, 38,* 603–610.

Skaalvik, E. M. (1997). Self-enhancing and self-defeating ego orientation: Relations with task and avoidance orientation, achievement, self-perceptions, and anxiety. *Journal of Educational Psychology, 89,* 71–81.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Stipek, D., & Mac Iver, D. (1989). Developmental change in children's assessment of intellectual competence. *Child Development, 60,* 521–538.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn and Bacon.

Turner, J. C., Midgley, C., Meyer, D. K., Gheen, M., Anderman, E., Kang, Y., & Patrick, H. (2002). The classroom environment and students' reports of avoidance strategies in mathematics: A multimethod study. *Journal of Educational Psychology, 94,* 88–106.

Urdan, T. (2004). Predictors of academic self-handicapping and achievement: Examining achievement goals, classroom goal structures, and culture. *Journal of Educational Psychology, 96,* 251–264.

Urdan, T., & Midgley, C. (2003). Changes in the perceived classroom goal structure and pattern of adaptive learning during early adolescence. *Contemporary Educational Psychology, 28,* 524–551.

Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology, 96,* 236–250.

(*Appendix follows*)

Appendix

Item Descriptions and Factor Loadings of Achievement Goal Scales

Item	Lower elementary	Middle elementary	Upper elementary	Middle school
Mastery-approach goal				
1. I like math work that I'll learn from, even if I make a lot of mistakes.	.49	—	.42	.64
2. An important reason I do my math work is because I like to learn new things.	.43	.58	.63	.75
3. I like math work best when it really makes me think.	.59	—	.60	.61
4. An important reason why I do my math work is because I want to get better at it.	—	.50	.54	.47
5. I do my math work because I'm interested in it.	.44	.47	.73	.70
6. I do my work in math because I want to learn as much as possible.	.50	.57	.71	.65
Performance-approach goal				
1. I would feel really good if I were the only one who could answer the teacher's questions in my math class.	—	—	—	—
2. I want to do better than other students in my math class.	—	—	.42	.59
3. I would feel successful in math if I did better than most of the other students in the class.	.55	.55	.53	.37
4. I'd like to show my teacher that I'm smarter than the other students in my math class.	.49	.47	.63	.67
5. Doing better than other students in math is important to me.	.49	.53	.62	.60
Performance-avoidance goal				
1. It is very important to me that I don't look stupid in my math class.	—	—	—	—
2. An important reason I do my math work is so that I won't embarrass myself.	.50	.68	.56	.64
3. The reason I do my math work is so the teacher doesn't think I know less than others.	.61	.66	.70	.71
4. The reason I do my math work is so others in the class won't think I'm dumb.	.61	.68	.74	.75
5. One reason I might not participate in math class is to avoid looking stupid.	.57	.64	.41	.49
6. One of my main goals in math is to avoid looking like I can't do my work.	.64	.66	.49	.66
Mastery-avoidance goal				
1. I'm afraid that I won't do my very best in my math class.	—	.44	.51	.50
2. I'm concerned that I may not learn all there is to learn from my math class.	—	.70	.73	.69
3. I'm afraid that I may not understand the lessons in my math class as completely as I should.	.54	.72	.70	.77
4. It is important to me "not" to do my math work incorrectly.	—	—	—	—
5. I worry that I may not learn all that I possibly could in math.	.70	.64	.71	.77
6. It is important to me to avoid the possibility of not learning in my math class.	.60	.72	.70	.41

Note. Results are from the final 2 × 2 CFA model (i.e., Model A) for each age group. Dashes indicate items removed from analyses because of factor loadings < .40. All remaining items formed the basis for all CFA models tested within each age group.

Received February 22, 2007
Revision received March 26, 2009
Accepted March 26, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 897–911

0022-0663/09/$12.00 DOI: 10.1037/a0015817

Pictures and Words: Spanish and English Vocabulary in Classrooms

Lee Branum-Martin, Paras D. Mehta, and David J. Francis
University of Houston

Barbara R. Foorman
Florida State University

Paul T. Cirino
University of Houston

Jon F. Miller
University of Wisconsin—Madison

Aquiles Iglesias
Temple University

The current study evaluated the relation between Spanish and English vocabulary. Whereas previously reported correlations have revealed strong differences among types of vocabulary measures used and the ages of the students tested, no prior study had used a multilevel model to control for classroom-level differences. The current study used multiple measures of vocabulary—picture vocabulary and narrative production tasks—in multilevel models of 1,300 Spanish-speaking students in 247 kindergarten and 1st-grade classrooms in English immersion and bilingual transitional programs. The current results highlight the need to separate classroom effects from student effects, since for vocabulary measures, student-level correlations were strongly biased toward zero when classroom-level correlations were opposite in direction from student-level correlations. Most important, the current results support a strong distinction between types of vocabulary measure (e.g., picture vs. narrative) and suggest sizable influence of instruction for questions of bilingual performance.

Keywords: bilingualism, vocabulary, multilevel models, narrative measures, instruction

Vocabulary is essential for understanding the meaning of spoken and written language. The more words known, the better prepared students will be for using and understanding languages (see reviews by Scarborough, 2001; Sénéchal, Ouellette, & Rodney, 2006). Because vocabulary is generally easy to measure, strongly related to reading comprehension, and frequently used as a proxy for general language competence in both English (Scarborough, 2001; Sénéchal et al., 2006; Wagner et al., 1997) and Spanish (August, Carlo, Dressler, & Snow, 2005; Carlo et al., 2004; Genesee & Geva, 2006; Snow & Kim, 2007), measuring vocabulary can aid our understanding of the cognitive and educational processes in language learning.

However, for bilingual students, the situation is much more complicated, as words can be known to varying degrees in either or both languages (Fernández, Pearson, Umbel, Oller, & Molinet-Molina, 1992; Oller, 2005; Oller & Eilers, 2002; Umbel, Pearson, Fernández, & Oller, 1992). Therefore, knowing the relation between Spanish and English vocabularies in bilingual populations might help researchers and educators to improve educational outcomes for these students (August et al., 2005; Carlo et al., 2004; Fitzgerald, 1995; Genesee & Geva, 2006; Snow & Kim, 2007). For example, is vocabulary competence language-specific (negatively correlated or zero), or does vocabulary competence generalize across languages? To what extent might instruction in one language aid performance in another language? Is vocabulary competence itself a single ability or a loose constellation of diverse subskills? Such questions require close examination of vocabulary measures and their correlation across languages.

Although there are numerous accounts in the literature regarding this correlation (e.g., Bialystok, Luk, & Kwan, 2005; Carlisle, Beeman, Davis, & Spharim, 1999; Gottardo, 2002; Lindsey, Manis, & Bailey, 2003; Nagy, García, Durgunoglu, & Hancin-Bhatt, 1993; San Francisco, Mo, Carlo, August, & Snow, 2006; for

Lee Branum-Martin, Paras D. Mehta, David J. Francis, and Paul T. Cirino, Texas Institute for Measurement, Evaluation, and Statistics, Department of Psychology, University of Houston; Barbara R. Foorman, Florida Center for Reading Research, Florida State University; Jon F. Miller, Department of Communicative Disorders, University of Wisconsin—Madison; Aquiles Iglesias, Department of Communication Sciences and Disorders, Temple University.

This work was supported in part by grants jointly funded by both the National Institute of Child Health and Human Development and the U.S. Department of Education's Institute of Education Sciences: "Oracy/Literacy Development of Spanish-Speaking Children" (No. HD39521) and "Biological and Behavioral Variation in the Language Development of Spanish-Speaking Children" (No. R305U010001), David J. Francis, principal investigator. The opinions expressed herein are those of the authors and do not necessarily reflect the attitudes and/or opinions of the funding agencies, and no endorsement of the findings is either granted or implied. We thank Jason Anthony for feedback on a draft of this manuscript and Karla Stuebing for helpful advice regarding the meta-analysis.

Correspondence concerning this article should be addressed to Lee Branum-Martin, Texas Institute for Measurement, Evaluation, and Statistics, Department of Psychology, University of Houston, Houston, TX 77204-6022. E-mail: Lee.Branum-Martin@times.uh.edu

an overview, see Genesee & Geva, 2006), the relation between vocabulary in Spanish and English varies on a number of factors, including the nature of vocabulary measures, student and sample characteristics, and the effect of instruction. Further, many of these studies have been limited to small sample sizes. Most important, however, none of these studies have taken into consideration the multilevel nature of the data, which is particularly relevant given that they were conducted in school contexts (Branum-Martin et al., 2006; Genesee, Geva, Dressler, & Kamil, 2006; Mehta, Foorman, Branum-Martin, & Taylor, 2005). In fact, the report of the National Literacy Panel on Language-Minority Children and Youth specifically called for more multilevel modeling of cross-language effects (Genesee et al., 2006). Therefore, the purpose of this work was to examine the relation between Spanish and English vocabulary—considering the relevant factors of measures, student characteristics, and instruction—and to conduct such an examination in a large sample in a multilevel context that takes into account the nested nature of the data.

What Is Known About the Relation Between Spanish and English Vocabulary?

Although theoretical discussions of bilingual abilities and research design cover many salient issues of measures, students, and instruction (e.g., Baker, 2001; Carroll, 1983, 1986; Cummins, 1979, 1983), there have been no systematic studies of the impact of these factors on the correlation between Spanish and English vocabulary. In general, the empirical findings are sparse for Spanish–English vocabulary (see Genesee & Geva, 2006; Snow & Kim, 2007). We searched the PsycINFO and ERIC databases, and the reference lists of the articles identified therein, for reported correlations between Spanish and English vocabulary measures for elementary school-age children. A total of 21 studies were found, which reported 32 correlation coefficients between Spanish and English vocabulary measures for children in prekindergarten through 6th grade. The correlation coefficients and their 95% confidence intervals are shown in Figure 1.

Figure 1 shows the 21 studies, grouped by test used and in order of grade level tested. Immediately to the right of the study name is an indication of the instruction or educational program in which students were taught (if reported). Each circle in the graph represents the level of the reported correlation, and the lines extend to the upper and lower bounds of the 95% confidence interval, calculated on the basis of the reported sample size. At the right-hand side of the figure are notations indicating the type of measure or test reported. Several studies reported more than one correlation and are therefore listed multiple times but with differences in the measure or instruction noted.

The first most striking observation about Figure 1 is that the groups of vocabulary measures (picture, experimental, or narrative) differ markedly in their Spanish–English correlation. Although there are a great number of methods for measuring vocabulary, including word lists, cloze procedures, and narratives (see Chapelle, 1998; Cronbach, 1942; Fiestas & Peña, 2004; Graves, 1987; Nagy & Scott, 2000; Nation, 2001; Pythian-Sence & Wagner, 2007; Read, 2000; Read & Chapelle, 2001; Webb, 2008), few of these have been reported in Spanish–English research. The measures included in Figure 1 are versions of the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 1997) and the Test de Vocabulario en Imágenes Peabody (TVIP; Dunn, Padilla, Lugo, & Dunn, 1986), versions of the Woodcock Language Proficiency Battery (WLPB; e.g., Woodcock, 1991; Woodcock & Muñoz-Sandoval, 1995), experimenter-constructed measures, and narrative-based measures. Overall, the correlations for the Peabody tests range from −.27 to .29, with a slight negative trend by student grade level, whereas the correlations for the Woodcock tests range from −.47 to .27, with a positive trend by grade. The four experimental measures are extremely varied in their level of correlation, and the correlations for the narrative measures are uniformly positive (r = .34 to .63).

Meta-Analysis

In order to substantiate these visual observations of Figure 1, we analyzed the correlations for the effects of grade and test type. The Fisher's Z-transform (Hedges & Olkin, 1985) of each of these correlations was predicted by grade level (prekindergarten–6th grade), test type, and the interaction of grade and test type. The model was fit in SAS PROC MIXED (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). In the analysis, the correlation for each study was weighted by its precision, the reciprocal of the standard error, so that better estimates counted more in the analysis. A multilevel model was examined to account for the nesting of correlations within-study, but there was not enough variability across studies, and estimates were not dependable, likely because of the small number of studies. Residual variances were allowed to differ across three test types: picture, experimental, and narrative. There were not enough combinations of grades by instruction types for dependable estimates of instructional effects (e.g., there were only four instances of mixed instruction and one of two-way instruction). Estimates from this model by test type and grade are shown in Table 1.

Table 1 shows the relative sizes and precision of the effects of grade and test type. The reference category for the tests was narratives, so the intercept of .506 is the Z value of narratives, with a main effect of .023 (*ns*) per year. The average correlation for both the PPVT (−.394) and the WLPB (−.831) was significantly lower than that for the narratives. Although the estimate for experimental measures was also negative (−.718), it was also very imprecise and therefore not significantly different from zero. The interactions of grade by test were not statistically significant—which was not surprising, given the small sample size—but may yield substantively important differences across grades (see Ziliak & McCloskey, 2008, for a discussion of practical vs. statistical significance). Therefore, these test effects and the interactions with grade were used to calculate model-based totals for each test type and grade and then transformed back to correlations. These model-predicted correlations are shown in Figure 2.

Figure 2 shows the model-predicted correlations by grade level for each test and helps to clarify visible trends in correlations from Figure 1. Each test has one line that extends from prekindergarten to Grade 6 (the range of grades reported in Figure 1). Each line has a marker for the midpoint of the grade levels represented by the studies. For example, the narratives represent only three grade levels from the three studies reporting the eight correlations (cf. Figure 1). The model results from Table 1 and Figure 2 help to confirm our observations of the

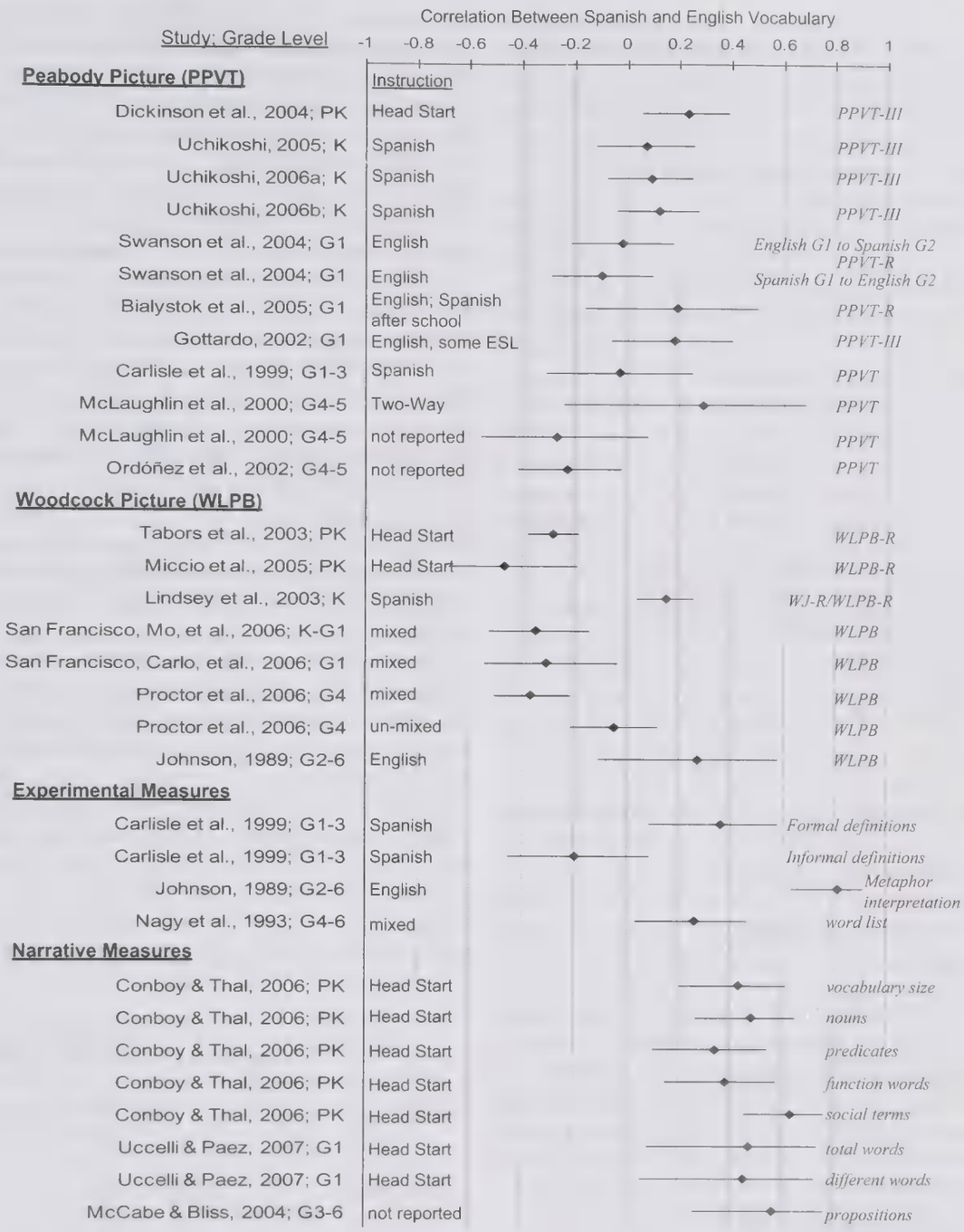


Figure 1. Correlation among prior studies between Spanish and English vocabulary measures, with 95% confidence intervals. Each study appears with the grade level of the sample (PK = prekindergarten; K = kindergarten; G = grade). Instruction type is shown immediately to the right of the vertical axis (ESL = English as a second language. Italics indicate special notes regarding the type of measure reported. "Un-mixed" instruction refers to one study (Proctor, August, Carlo, & Snow, 2006) in which language of instruction was used as a predictor, statistically controlling for Spanish instruction. PPVT = Peabody Picture Vocabulary Test (R = Revised; III = Third Edition); WLPB-R = Woodcock Language Proficiency Battery (R = Revised); WJ-R = Woodcock–Johnson Psychoeducational Battery—Revised Tests of Achievement.

prior studies presented in Figure 1. Narratives have a consistently high positive correlation between Spanish and English, with no strong trend across grades. Experimental tests have a wide-ranging set of correlations, which are unlikely to be dependable, given that each of the four tests was a unique measure of vocabulary. The helpful clarification from Figure 2 is that the differing grade trends of the PPVT and Woodcock tests are shown. Although the individual estimates for the effects and interactions of grade might not be statistically significant, the resulting model-based correlations imply substantively different correlations, which clarify our reading of the prior studies in Figure 1. Together, Figures 1–2 and Table 1 show influences on Spanish–English vocabulary correlations from the nature of the measures, student characteristics, and

Table 1
Model Results for the Meta-Analysis of the Correlation Between Spanish and English Vocabulary From Figure 1

Effect	Results	
	Estimate (*Z*)	*SE*
Intercept (narrative)	0.506*	0.042
Grade (narrative)	0.023	0.021
PPVT/TVIP	−0.394*	0.100
WLPB	−0.831*	0.112
Experimental	−0.718	0.861
Grade × PPVT	−0.050	0.039
Grade × WLPB	0.067	0.045
Grade × Experimental	0.183	0.250
	Variance	*SE*
Residual variances		
Picture tests	0.006*	0.002
Experimental	0.056	0.056
Narrative	0.002*	0.001

Note. Regression estimates are in the metric of Fisher's *Z*. The residual variances for the different test types were not homogeneous, $\chi^2(2) = 12.3$, $p = .002$. Model-predicted correlations are shown in Figure 2. PPVT = Peabody Picture Vocabulary Test; TVIP = Test de Vocabulario en Imágenes Peabody; WLPB = Woodcock Language Proficiency Battery (including the revised version).
*Parameter is significantly different from zero, $p < .05$.

classroom contexts. These results from the prior literature also highlight areas in which we have little or no knowledge regarding these three areas of potential influence.

Nature of Vocabulary Measures: Format, Method, and Content

The first important finding regarding prior studies of Spanish–English vocabulary is that the nature of the measure is a major source of variability in the obtained Spanish–English correlation. In Figure 1, the blocks of tests (PPVT, WLPB, experimental, and narratives) each yielded substantively different sets of correlations, as emphasized in Figure 2 and the significant test effects in Table 1. One reason for this empirical distinction may derive from the theoretical distinction often made between receptive and expressive vocabulary. Receptive vocabulary represents the words that can be understood, whereas expressive vocabulary represents the words that can be used in speech or writing (Henricksen, 1999; Nation, 2001). In terms of format, receptive measures typically require pointing, marking, or other nonverbal responses from among a limited response set, whereas expressive measures typically require the participant to structure an open-ended oral response. The PPVT and the TVIP are receptive measures requiring the student to choose a picture that represents the word spoken by the examiner, for example, "Which picture shows *running*?" The WLPB Picture Vocabulary subtest is an expressive measure that requires the student to verbally name a picture, in response to a question of "What is this?" (after being presented with up to seven receptive items at the beginning of the test; this subtest is usually administered only to children at the kindergarten level or below).

Expressive measures typically involve verbal abilities not utilized for receptive measures, such as articulation, syntax, fluency, and aspects of social communication. Expressive measures themselves also may vary in terms of the type of response required, ranging from, for example, a naming type of task (as in the WLPB Picture Vocabulary subtest) to a narrative task, such as the telling of a story, where vocabulary is assessed in the context of more naturalistic language (e.g., Conboy & Thal, 2006; Miller et al., 2006; Miller & Iglesias, 2003–2004; Tilstra & McMaster, 2007; Uccelli & Paez, 2007). In narratives, several measures can be gathered from a single piece of discourse, such as the number of different words and clauses, grammatical complexity, and utterance length. This range of possible measures may allow narratives to act as a more valid index of the size of vocabulary, especially as used in actual speaking contexts (see Miller et al., 2006; Miller & Iglesias, 2003–2004; Tilstra & McMaster, 2007). Empirically, there is evidence to suggest that expressive and receptive measures have different relations to measures of comprehension in English, emphasizing the distinction between these as skills (Ouellette, 2006; Wise, Sevcik, Morris, Lovett, & Wolf, 2007). Given the differences between receptive, expressive, and narrative methods, it is not surprising that Spanish–English correlations differ among the PPVT, the Woodcock, and narrative measures shown in Figure 1.

The content of an individual measure may also influence the relation between Spanish and English vocabulary. For example, the degree to which respective Spanish and English measures include cognates, such as bicycle and *bicicleta,* would contribute to high cross-language correlations (Carlo, 2001; Carlo et al., 2004; Nagy et al., 1993; Umbel et al., 1992). Alternatively, a test could specifically avoid such similarities or even use false cognates (e.g., *embarazada* means "pregnant," not "embarrassed"). Further, the concepts of one test could match across languages (e.g., words for "boat" could be tested in both languages), or the test could be constructed to avoid similar concepts or even to have

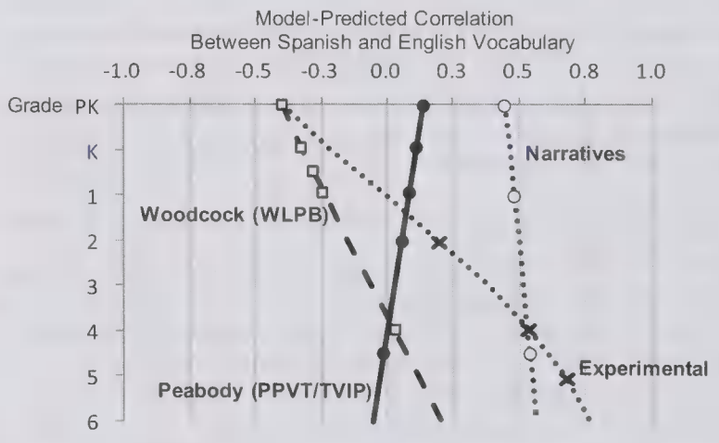


Figure 2. Grade trends for model-predicted correlations between Spanish and English vocabulary from the analysis of prior studies. Fisher's *Z*-transforms of the correlations in Figure 1 were analyzed in a regression model in which the Z-transformed score was predicted by test type (Peabody Picture Vocabulary Test / Test de Vocabulario en Imágenes Peabody [PPVT/TVIP] Woodcock Language Proficiency Battery [WLPB], experimental, or narratives), grade level (PK = prekindergarten; K = kindergarten), and the interaction of grade and test type. The regression model results (see Table 1) were transformed back to correlations and graphed by grade level here. Markers on the lines represent the midpoints of grade levels represented by studies. For details of the model, see the text.

no overlap whatsoever in the concepts tested. Thus, measures with higher conceptual similarity across languages would tend to produce higher cross-language correlations, as the similarity might aid recognition or production of word meanings in both languages (Dressler & Kamil, 2006). Both the Woodcock and the Peabody tests are nearly equal in their proportion of cognates, although the Woodcock test contains more matched concepts across Spanish and English (e.g., cow/*vaca*). In the PPVT-III, 35 of 125 Spanish words (28%) have the same picture among the 204 English words (note that the PPVT contains more items). In the WLPB–R tests, 41 of the 58 words (85%) in Spanish have the same pictures as in English. Thus, the Woodcock test has more items that are semantically the same across languages than does the Peabody test. Despite this cross-language similarity of concepts on the WLPB, the correlations for the WLPB were lower than for the PPVT, $F(1, 22) = 8.3$, $p = .008$, as analyzed from the estimates in Table 1.

With regard to the nature of the measures, it can be seen in Figures 1–2 that narrative measures are uniformly positively related across language, whereas the expressive Woodcock measures are typically negatively related across language and the receptive Peabody measures range from slightly positive to slightly negative, depending on age. These results highlight the fact that because narrative measures draw on expressive speaking skills, these skills may be highly related across language. Whereas the Woodcock is an expressive task, the scoring is relatively stringent, ensuring that specific word knowledge is tested. The receptive Peabody is generally more positively related across language than is the Woodcock. Despite the prevalence of across-language concepts and the expressive nature of the test, the Spanish–English correlations for the WLPB are, in general, lower than those of the PPVT (see Figure 1 and Table 1).

Student and Sample Characteristics

Beyond the differences in measures, Figures 1–2 show some evidence of age trends: In the Woodcock tests, older children have more positive relations between Spanish and English vocabulary, whereas a slightly negative age trend is present in the Peabody tests. Indeed, individual characteristics, such as age, native language, motivation, and other linguistic and intellectual abilities, may be important influences on across-language vocabulary correlations (see discussion by Bachman & Palmer, 1996; Dressler & Kamil, 2006). For example, given that vocabulary size grows rapidly with age (Pythian-Sence & Wagner, 2007), it could be conjectured that younger children simply do not know as many words, and the potential for the smaller number of words to be known across two languages will be lower (resulting in lower cross-language relations). Figures 1–2 show that such an age trend may be plausible for the Woodcock and experimental measures, but for the Peabody studies, a negative trend over age in the correlations can be seen: the older the students, the more negative is the relation between Spanish and English vocabulary. Although there may be age trends, these seem to be dependent on the vocabulary measure used.

Native language of a given sample will also influence the relation of Spanish and English vocabulary. There are likely to be a host of differences between native English speakers and native Spanish speakers along social, economic, and instructional lines, and therefore differences may be found between these samples (August & Hakuta, 1997; Bachman & Palmer, 1996). Although it is possible that fully bilingual persons could constitute another group or blur the "native" distinction altogether, the very correlation we seek is itself a measure of how "bilingual" the group is—the more highly correlated the measures of vocabulary, the more unified is their word knowledge. The majority of studies shown in Figure 1 sampled low-socioeconomic status (low-SES) Hispanic students. Although a few studies included exceptional students (e.g., fully biliterate or disabled), there are not enough such studies to draw distinctions among these characteristics across all combinations of measures and ages of the students.

We identified only one study evaluating native English speakers in a program in which Spanish- and English-speaking students were taught in both languages, with the goal of each group achieving literacy in both languages (McLaughlin, August, & Snow, 2000). This correlation for a mixed sample of students was markedly higher ($r = .29$), whereas all of the other correlations for this age group tended to be negative. This could suggest that the learning processes for non-native Spanish speakers are substantially different than they are for samples of only native Spanish speakers, but clearly more such studies would be needed to establish this. More importantly, however, if we were interested in differences due to native language, the Spanish–English correlation for each student language group would be reported separately. However, this one study could also suggest that the instructional program involving mixed student backgrounds is different, a possibility that is discussed in the next section.

Finally, the environmental or sociocultural context that a given student experiences is also likely to influence the relation of Spanish and English vocabulary. These characteristics could include generational status (e.g., foreign-born, first or second generation), the availability of both languages within the community, and the extent to which either or both languages are spoken or heard within home environments (e.g., through family members or mass media). These types of contextual variables rarely have been systematically sampled and are generally absent from Figure 1. Such contextual variables are the focus of the next section.

Context Effects: Instruction and Classroom Differences

Figure 1 lists brief indications of the type of instruction for each study, if reported. Within the picture vocabulary studies in the top half of Figure 1, the correlations are quite similar across Spanish and English instruction. However, because there were only 1 to 8 coefficients each for Head Start, Spanish, English, mixed, and not reported groups, we could not obtain dependable model estimates for types of instruction. In the given correlations, there do not appear to be any obvious program effects beyond the effects of the measures and the ages of the students.

It is unfortunate that the literature has little to offer on the effects of types of instruction across ages and measures, because the Spanish–English instructional program or philosophy could be argued to be a major source of classroom differences. There are many types of instructional program, each with different goals, including English immersion, transitional, maintenance, and dual language (for an overview, see Baker, 2001). In immersion, stu-

dents are taught in a single language (usually the majority language of the culture). In transitional education, students are instructed at least partially in their native language and, as they grow more proficient, more of the majority language is used in instruction. Maintenance programs aim to use enough of the students' native language to preserve their linguistic proficiency while promoting acquisition of the majority language. Whereas native or foreign language programs are often chosen for an entire school, sometimes within schools, classrooms may have different instructional programs. With differing goals of educational programs, classrooms may therefore differ greatly in the type and amount of instruction for particular languages. For example, English immersion for Spanish-speaking children might raise mean performance in English but would likely lower Spanish–English correlations, whereas other programs using more Spanish instruction might produce higher Spanish outcomes and higher Spanish–English correlations. Such hypotheses of a trade-off can be examined as effects of classroom program.

Even within the same bilingual program, however, teachers may spend their time in different ways, further contributing to how classrooms may differ from each other. Thus, whereas teachers in an immersion program, for example, would readily be expected to differ from teachers in a transitional program, those immersion teachers may spend their instructional time very differently from each other (e.g., Cirino, Pollard-Durodola, Foorman, Carlson, & Francis, 2007; Saunders, Foorman, & Carlson, 2006). For example, some immersion teachers may spend more time on vocabulary instruction than other immersion teachers, who might emphasize spelling or other tasks. Thus, classrooms likely differ in the amount (and perhaps quality) of instruction in various tasks. Instructional time spent in vocabulary tasks in both languages could be expected to increase Spanish–English correlations, whereas more language-specific instruction or nonvocabulary instruction might produce lower Spanish–English correlations. If measures of instructional time and language are available, then there is the potential to go beyond mere nominal program distinctions (as noted in Thomas, 1992) toward more detailed measures of how classrooms may differ from each other.

In summary, the empirical results on the correlation between Spanish and English vocabulary measures show a strong influence for the type of measure chosen (picture, experimental, or narrative). There are some possible effects of age, which differ by measure, but this is likely contaminated by instructional effects, since in the United States most students are moved toward more English instruction as they grow older. Other student characteristics, such as economic status or general language competence, were not reported or sampled widely enough to estimate effects. There are no clearly discernible general effects of instruction (e.g., Spanish vs. English), largely because not enough studies have been reported across the different measures and grades. In short, the most important things we know about prior studies are that the measurement procedure matters and that grade level may also matter for inferences regarding the relation between Spanish and English vocabulary. For picture tests, the Spanish–English correlations are low to negative, suggesting unrelated student abilities, whereas the correlations for narrative measures are positive, suggesting similar student abilities across languages.

Taking It to the Next Level: Classrooms Are Different and Instruction Matters

An important limitation to all prior reported studies in Figure 1 is that classrooms may differ in important ways that were not modeled. Students spend much of their time in classrooms, which may differ systematically from one another, even within the same school. Because instruction is typically delivered to students as groups in classrooms, if classrooms differ in average performance, then this would be an important source of variability to consider in examining literacy outcomes (Mehta et al., 2005). Wherever classrooms exhibit systematic differences, empirical models can be improved by modeling these classroom differences. Thus, examining the correlation between Spanish and English vocabulary becomes an examination of two correlations: one at the classroom level and another at the student level.

Multilevel models are gaining recognition in education as a helpful way to separate classroom-level effects from student-level effects (Mehta et al., 2005; Mehta & Neale, 2005). In a multilevel model for students within classrooms, we separate the variability in the outcome (e.g., Spanish vocabulary) into a classroom component and a student component. Such single-outcome (univariate) multilevel models are common in education and allow researchers to examine student-level influences on the outcome separately from classroom-level influences. In a similar way, we can take two outcomes (e.g., Spanish and English vocabulary) and examine the classroom-level variance and student-level variance. Additionally, we can examine the covariance (and correlation) between these two outcomes, both at the classroom level and at the student level.

The substantive impact of this separation is that we now have a correlation at the classroom level (effectively between classroom means) and also a correlation at the student level, after controlling for those classroom-level differences. This separation is important, not only because it models differences we believe to be important but also because it allows our substantive questions to be examined at the appropriate level: classroom influences that occur at the classroom level (e.g., programs, instruction, and school–community contexts) and student influences (e.g., age, language, literacy, or other personal characteristics) that occur within—after controlling for—those classroom influences. Technically, such a model is a multilevel multivariate random intercepts model or a multilevel structural equation model (Curran, 2003; du Toit & du Toit, 2003; Goldstein & McDonald, 1988; Kaplan & Elliot, 1997; Mehta & Neale, 2005; B. O. Muthén, 1991, 1994; Rabe-Hesketh, Skrondal, & Pickles, 2004). Multivariate, multilevel studies have been specifically called for in the area of cross-language relations by the National Literacy Panel on Language-Minority Children and Youth (Genesee et al., 2006).

The reason this substantive distinction between classroom and student effects is so important for bilingual education is that the Spanish–English correlation we seek may be utterly different for classrooms than it is for students. In fact, the classroom-level correlation may be opposite in direction to that of the student-level correlation (c.f., Snijders & Bosker, 1999). For example, classroom instruction may be quite specialized in one language at the expense of the other (negative relation), whereas students tend to perform similarly in both languages (positive relation). Ignoring the distinction between classrooms and students in such a case may

produce an overall Spanish–English correlation around zero—leading us to erroneously infer that student abilities are unrelated.

It is worth noting that there are other types of context effects, such as differences among schools and communities. However, such differences have not been systematically sampled in the prior literature. Such differences can be modeled as effects at the classroom level or can be fully expanded to a third level (Goldstein, 2003; Raudenbush & Bryk, 2001).

The above review highlights several deficiencies in the current literature. First, none of the studies in Figure 1 distinguished classroom-level effects from student-level effects. The ways in which such classrooms may differ include the instructional model utilized and the extent to which English or Spanish is actually used in the classroom, regardless of the nominal program label. Second, very few studies consider multiple types of measures within the same population, which is important given that different patterns of correlations could imply that different constructs are being measured. That is, construct validity may differ in a multitrait–multimethod sense (Campbell & Fiske, 1959; Eid, Lischetzke, & Nussbeck, 2006). Third, sample sizes are typically small. Therefore, the purpose of the present study was to examine the relation between Spanish and English expressive vocabulary in a large sample of native Spanish-speaking students, considering two types each of vocabulary measures, instructional programs, and language usage, and to conduct such an examination with a multilevel model of students within classrooms.

Method

The following research questions were asked in the current analysis:

1. What is the student-level correlation between Spanish and English vocabulary as measured by picture vocabulary word tests, and as measured by a narrative task?
2. What is the classroom-level correlation between Spanish and English vocabulary as measured by picture vocabulary word tests and as measured by a narrative task?
3. What do the cross-method, cross-language correlations at each level suggest about the convergent and discriminant validity of the tasks?
4. How strongly does the balance of instructional English used in the classroom relate to these vocabulary outcomes?

Participants

The current sample included 1,300 students who were native Spanish speakers in 247 classrooms in 32 schools in border Texas, urban Texas, and urban California. The sample was taken from end-of-year (April–May) testing in kindergarten and first-grade classrooms from a larger longitudinal project (Francis et al., 2005). The schools were selected to have been academically acceptable by their respective states but were generally located in low-SES, predominantly Hispanic neighborhoods (Branum-Martin et al., 2006; Francis et al., 2005). A total of 81% of the parents were foreign-born, and 79% of the students came from households earning less than $30,000 per year. The number of students and classrooms for each grade and program is shown in Table 2, with mean age and percentage of female participants. The students were balanced on gender and age, as was expected for end-of-year measures.

Table 2
Description of the Sample

Grade/program	No. of students	No. of classrooms	Mean age in years (*SD*)	% female
Kindergarten				
Immersion	351	35	6.1 (0.3)	52%
Transitional	606	50	6.2 (0.4)	51%
First grade				
Immersion	474	75	7.1 (0.4)	53%
Transitional	625	80	7.2 (0.4)	52%
Total unique	1,300	247		

Note. Total unique refers to the number of individual students or instructors across the two grades.

Measures and Procedures

Picture vocabulary. Two types of vocabulary measures were used in each language: picture vocabulary and narrative elicitation. The Picture Vocabulary subtest is an expressive picture-naming task that was taken from the Woodcock Language Proficiency Battery—Revised (WLPB–R; Woodcock, 1991; Woodcock & Muñoz-Sandoval, 1995). Students were asked by bilingual examiners to name the picture shown. Credit was given only for correct answers in the language of the test. For all students, Spanish tests were given first, with English tests generally completed within 1–2 weeks following the Spanish test. W-scores were used in order to have scores that are comparable across grades on a single metric.

Narrative elicitation: Number of different words. In the narrative elicitation method, children were individually presented with one of Mercer Mayer's wordless *Frog* picture books and asked to tell a story to the pictures. The examiner told the child a story in either English or Spanish and the child was to then retell the story in that same language. Each child's narrative was recorded and later transcribed and analyzed with Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2003–2004). SALT is a set of procedures for transcription, combined with a computer program for analyzing transcripts for number of words, sentences, and other linguistic features. For the current study, the number of different root words used by the child was taken as an index of vocabulary size. For further details of the procedure, see Miller et al. (2006).

Percentage of instructional English. In the current analysis, we selected teachers who designated themselves as following an immersion (English-only) or a transitional education program. The goal of the transitional program was to use Spanish during instruction, with a gradual shift over time toward more English, with mainstream English-only instruction occurring by the end of second grade. In the current sample, no transitional classrooms had officially switched to English-only instruction.

Because time and budget constraints prohibited observation of every classroom, a random subsample of 31–46 classrooms in each grade and program (kindergarten immersion = 31; kindergarten transitional = 36; first-grade immersion and transitional = 46 each) was observed twice per year using a protocol in which

each minute of instruction was coded for language and content (for full details, see Foorman, Goldenberg, Carlson, Saunders, & Pollard-Durodola, 2004). Observers were trained twice per year to achieve interrater reliability above 80%, with in-classroom checks by the coordinator on 10% of the observations. For the purpose of the current study, data from each observation yielded number of minutes of English instruction, number of minutes of Spanish (or mixed-language) instruction, and total number of minutes (which included noninstruction, such as lining up or giving directions). For simplicity, we ignore the content codes (e.g., spelling vs. phonology) to focus on an overall effect of Spanish or English language used for instruction.

For each child, there were records of the teachers who served that classroom in addition to the primary reading and language arts teacher (e.g., some schools had additional English language development instructors). Because several schools shared teachers among classrooms, a ratio was constructed for each teacher to represent the percentage of time in which English was used for instruction versus the amount of time Spanish was used (i.e., time in English divided by total observed instructional time). We therefore used this percentage to reflect the balance of instructional English language used in the classroom, ranging from 0 to 1.0, equivalent to 100% Spanish to 0% Spanish.

Table 3 presents descriptive statistics for the percentage of English instruction by grade and program. The first line of Table 3 shows that teachers in the 31 observed kindergarten immersion classrooms spent, on average, 92% of their time speaking English, compared with only 24% of time in kindergarten transitional classrooms. The descriptive statistics show that, on average, teachers were faithful to their program labels, although there was considerable variability within each program. For example, some English immersion classrooms were observed to spend as little as 13–49% of their instructional time in English.

Analysis

We used two modelsto answer the four research questions in the current study. The first model was an unconditional multivariate random intercepts model, used to estimate the student- and classroom-level covariance structures. The second, conditional model was the same model, but with the classroom-level observed percentage of instructional English predicting the four classroom-level outcomes (for examples of such models for students within classrooms, see Branum-Martin et al., 2006; Mehta et al., 2005; Mehta & Neale, 2005).

Table 3
Classroom-Level Percentage English Instruction by Grade and Program

Grade/program	No. classrooms	*M*	*SD*	Minimum	Maximum
Kindergarten					
Immersion	31	92.4	14.3	49.5	100.0
Transitional	36	24.2	18.4	0.7	99.6
First grade					
Immersion	46	84.3	20.3	13.9	99.4
Transitional	46	29.4	24.4	1.6	96.7

Whereas these models sound complicated, they are easy to conceptualize. With two types of measures in two languages (picture vocabulary and narratives in both Spanish and English), there were four outcomes to consider. Because a multilevel analysis separates the outcomes into two levels, we could consider a 4×4 correlation matrix among these outcomes both for students and for classrooms. The cross-language correlations are presented graphically, as in Figure 1, and each two-level 4×4 matrix is also presented.

Each of the two models was fit separately for kindergarten and first grade. The two programs—immersion and transitional—were designated as separate models in a multiple-group analysis. The regressions of each classroom-level outcome on instruction were held equal across the two instructional groups so that instructional effects were consistent—that is, that a given amount of instruction would have the same effect in both kinds of program. The models were fit with Mplus 5 (L. K. Muthén & Muthén, 2007).

Results

The first models included four variables (picture vocabulary and number of different words, for both Spanish and English) with random intercepts for classrooms. Each was a two-group model (immersion and transitional), with variances, covariances, and intercepts free to differ between groups. Table 4 shows the standardized results (correlations) for each group. The correlation matrices show classroom correlations above the diagonal and student correlations below the diagonal. Whereas the statistically significant correlations are designated with asterisks, it is important to realize that the classroom level estimates are based on 31–46 classrooms and thus have wide confidence intervals but should not necessarily be dismissed as substantively zero (cf., Ziliak & McCloskey, 2008).

Picture Vocabulary

The corners of the matrices in Table 4 show the Spanish–English correlation for each task. In kindergarten immersion, classroom averages on the WLPB Picture Vocabulary subtest were correlated −.76 (upper top left), whereas student WLPB Picture Vocabulary performance was correlated .05 (lower top left). For the kindergarten transitional programs, WLPB Picture Vocabulary was correlated −.41 (upper top left) at the classroom level and .05 (lower top left) at the student level. These classroom-level correlations for picture vocabulary are strongly negative, whereas the student-level correlations are near zero. These results provide a sharp contrast to the results of previous studies presented in Figures 1–2, where student-level correlations were negative for kindergarten and first grade.

Narrative Number of Different Words

The same correlations for the number of different words in the narrative task have quite a different pattern, shown at the bottom right of each matrix. For kindergarten immersion, the Spanish–English correlation is .64 at the classroom level and .48 at the student level. These same kindergarten correlations in the transitional group are −.13 and .42 at the classroom and student levels, respectively. Examining these correlations for first grade, we see

Table 4
Multilevel Correlation Matrices

Variable	Kindergarten				First grade			
	PV		NDW		PV		NDW	
	Spanish	English	Spanish	English	Spanish	English	Spanish	English
	Correlations							
Immersion								
PV								
Spanish	—	−.76*	.41	−.43	—	−.97*	.54	−.73*
English	.05	—	.04	.73*	.03	—	−.49	.71*
NDW								
Spanish	.36*	.02	—	.64	.38*	.03	—	.06
English	.08	.44*	.48*	—	.04	.44*	.51*	—
Transitional								
PV								
Spanish	—	−.41	.56	−.07	—	−.88*	.34	−.59*
English	.05	—	−.51	.88*	.14*	—	−.25	.70*
NDW								
Spanish	.27*	.08	—	−.13	.35*	.13	—	.36
English	.06	.62*	.42*	—	.10	.60*	.45*	—
	Intercepts and standard deviations							
Immersion								
Intercept	431	451	63	69	444	458	66	79
Classroom *SD*	8	7	7	10	13	10	9	10
Student *SD*	19	14	19	21	22	15	18	22
ICC	.15	.21	.11	.18	.25	.28	.19	.18
Transitional								
Intercept	462	430	73	55	469	444	75	67
Classroom *SD*	8	6	8	8	12	7	8	9
Student *SD*	20	15	18	24	22	17	17	22
ICC	.14	.13	.17	.11	.24	.15	.19	.14

Note. Two multivariate random intercepts models were fit with immersion and transitional groups, separately for kindergarten and first grade. PV = Picture Vocabulary subtest W score from the Woodcock Language Proficiency Battery—Revised; NDW = number of different words from narrative story; ICC = intraclass correlation. Correlations above the diagonal represent those between classrooms; correlations below the diagonal represent those within classrooms (students). Each model was fully saturated. Kindergarten: log likelihood = −14139; Akaike information criterion (AIC) = 28,374 for 48 parameters. First grade: log likelihood = −16243; AIC = 32,582 for 48 parameters. The grade-referenced W scores for kindergarten and first grade were 463 and 474, respectively.
* $p < .05$.

that the student level appears similar, between .45 and .51, whereas the classroom correlation is .06 for immersion and .36 for transitional.

Across-Measure Correlations

The off-diagonal elements provide a rich source of information not available in Figure 1: the comparison of the different tasks within and across languages. At the classroom level, Spanish picture vocabulary was correlated positively with Spanish narrative number of different words (.34 to .56), and English picture vocabulary was highly correlated with English narrative number of different words (.70 to .88). These suggest that classrooms with high-average performance on picture vocabulary also performed highly on the narrative vocabulary task in the same language. The across-language and across-task correlations suggest virtually no relation between classroom performance on Spanish narrative number of different words with English picture vocabulary (.02 to .13). The classroom relations between English narrative number of different words and Spanish picture vocabulary are all negative but much more variable: −.43 and −.73 in kindergarten and first-grade immersion, and −.07 and −.59 for kindergarten and first-grade transitional.

At the student level, these off-diagonal correlations appear fairly homogeneous. The correlations between Spanish narrative number of different words and Spanish picture vocabulary ranged between .27 and .38. The student correlations between English narrative number of different words and English picture vocabulary were .44 for immersion and .60 to .62 for transitional. The student-level correlations between Spanish picture vocabulary and English narrative number of different words ranged between .04 and .10. As with the classroom level, the student-level correlations between English picture vocabulary and Spanish narrative number of different words are more variable but mostly negative, ranging from −.51 to .04.

The intercepts and standard deviations are shown at the bottom of Table 4. The W-score metric is a developmental scale that is comparable across grade levels (Woodcock, 1991). The narrative number of different words provides a similarly comparable measure, because this value represents the number of different root words used. The language imbalance of the instructional groups

can be seen in that the English immersion group has higher English performance, whereas the transitional group that received more Spanish instruction has higher performance in Spanish. For reference, the W-score of 431 corresponds to an age equivalent of 2 years, 10 months; 444 corresponds to 3 years, 5 months; 458 corresponds to 4 years, 5 months; and 469 corresponds to an age equivalent of 5 years, 7 months.

The intraclass correlations (ICCs) are shown at the bottom of Table 4 for each measure in each grade and program. The ICC represents the proportion of variability due to classrooms—or the extent to which classrooms differ. In this way, the ICC indexes the need for multilevel models to adequately separate differences into their appropriate levels. The ICC values presented in Table 4 ranged from .11 to .28, showing that up to 28% of the variability in the outcomes is related to simply knowing the classroom in which the student was located.

The only visually striking difference is that the ICC for picture vocabulary in English appears to be higher for immersion than for transitional. This suggests proportionately wider classroom differences on this measure within the immersion program than within the transitional program.

Effects of English Instructional Language

In order to examine the effect of instruction on classroom performance, the four outcomes (Spanish and English picture vocabulary and narrative number of different words) were regressed on the percentage of instructional English. These regression parameters are shown in Table 5. Table 5 presents the regression intercept for each program, the regression parameter (b) and its standard error, the standardized regression coefficient (β), and R^2 for each program. As noted in the Method section, the regression parameters were held equal between the immersion and transitional groups.

The intercept columns show the expected level of vocabulary performance for 100% Spanish instruction. For reference, these can be compared with the intercepts at the bottom of Table 4. The first line of Table 5 shows that in kindergarten, for each percentage point of English instruction, a decrease of .23 W-score units would be expected in Spanish picture vocabulary. The standardized coefficients show that the correlations between English instruction and Spanish PV were −.41 for immersion and −.52 for transitional. Overall, Table 5 shows that English instruction is negatively related to Spanish outcomes and positively related to English outcomes.

Despite the small number of classrooms, some of the regressions were statistically significant (kindergarten English picture vocabulary in transitional programs, $r = .67$; both picture vocabulary variables in both programs in first grade). These results suggest that instructional time in a particular language positively affected vocabulary performance in that language. To assist evaluation of these regressions, we also present R^2 values in Table 5. These R^2 values (.02 to .53) show that percentage of instruction was related to up to 53% of the variance in classroom-level vocabulary performance, albeit with much variability.

Comparison With Previous Studies

The correlations in Table 4, however, are difficult to place in the context of the prior results shown in Figure 1. Therefore, the cross-language correlations for each measure, picture vocabulary and narrative, were taken from the table and put into a graph in Figure 3. Figure 3 shows the Spanish–English correlations for the picture vocabulary measures and narrative measures at both the classroom and student levels. For each grade and program, there are two correlations connected by a dashed line: classroom and student correlation between Spanish and English. The 95% confidence intervals are calculated in the same manner as in Figure 1 and are based on sample size (classroom and student, as appropriate). The top half of Figure 3 shows the correlations from Table 4 for the picture vocabulary measures in the current study. In addition, ignoring the multilevel nature of the data, we present unilevel

Table 5
Regressions of Classroom Outcomes on Percentage of English Instruction

	Intercept				β		R^2	
Grade/outcome	Immersion	Transitional	b	SE	Immersion	Transitional	Immersion	Transitional
Kindergarten								
PV								
Spanish	454	468	−.24	.12	−.41	−.52	.16	.27
English	432	425	.20	.13	.37	.67	.14	.45
NDW								
Spanish	68	75	−.06	.14	−.13	−.14	.02	.02
English	40	48	.31	.22	.43	.73	.18	.53
First grade								
PV								
Spanish	467	481	−.28*	.08	−.49	−.66	.24	.44
English	440	436	.20*	.05	.48	.64	.23	.42
NDW								
Spanish	74	80	−.98	.11	−.25	−.31	.06	.10
English	69	63	.11	.11	.23	.33	.05	.11

Note. PV = Picture Vocabulary subtest W score from the Woodcock Language Proficiency Battery—Revised; NDW = Number of different words from narrative story.
* $p < .05$.

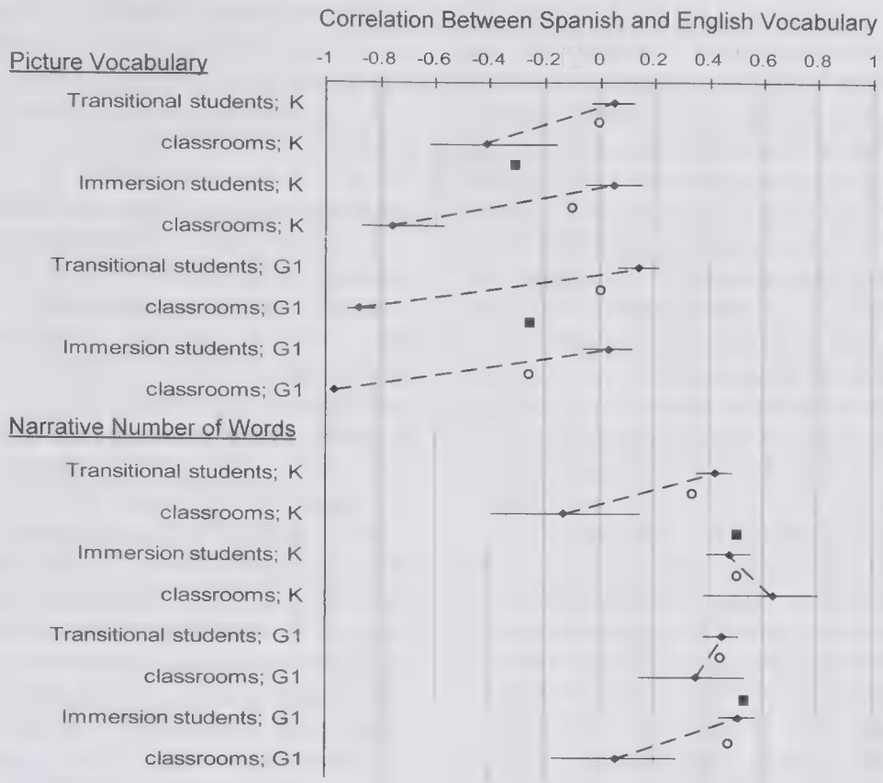


Figure 3. Current study multilevel correlations between Spanish and English vocabulary measures, with 95% confidence intervals. K= kindergarten; G1 = first grade. The correlations shown here are reproduced from Table 4. For each program and grade, two correlation estimates are shown: first, the student level, then the classroom level. Each student–classroom pair of estimates is joined by a dashed line. For reference, the empty circles represent the estimate obtained from a unilevel correlation, disregarding classroom differences. Solid squares show the expected correlations by grade and measure from the meta-analysis of prior correlations, which also disregarded classroom differences (see Figure 2).

correlations for each grade and program as open circles on the graph. Asterisks in Figure 3 show the model-expected correlations by measure and grade from the meta-analysis (calculated from Table 1 and shown in Figure 2).

The top half of Figure 3 shows that the student-level correlations for picture vocabulary measures are uniformly low but positive, whereas the classroom-level correlations are strongly negative. The lower half of Figure 3 shows the Spanish–English correlations for the narrative measures. The top portion of each pair shows that student-level estimates are uniformly positive (r = .42–.51), whereas the bottom portion of each correlation pair shows that classroom level Spanish–English correlations were far more variable, ranging from −.13 to .64.

Discussion

The current study is the first to examine multiple measures of vocabulary knowledge in a multilevel model in order to separate issues of constructs, student characteristics, and instructional language at their appropriate levels for two different programs of instruction: English immersion and transitional education. The current analysis, taken together with the systematic review of prior studies shown in Figures 1–3, yields conclusions in three major areas for bilingual vocabulary research: classroom differences, instruction, and measurement. Each of these areas is discussed in turn.

Classroom Differences: Multilevel

First, unacknowledged classroom differences can be dangerous for thinking about bilingual relations. In the current study, the classroom-level correlations for picture vocabulary were generally opposite in direction from those at the student level, implying that simple, unilevel correlations such as those reported in prior literature (Figure 1) would be biased toward zero. Such bias might tempt researchers to erroneously infer that student abilities are less positively related than they truly are. The unilevel correlations from the current study were similar to those of prior studies (see open circles versus asterisks in Figure 3), suggesting that previous studies may have suffered from this bias as a result of negative relations at the classroom level.

Instruction: Programs and Time

The second major finding was that instruction makes a difference. The dearth of multiple studies across ages and measures suggests that there is much more we need to know about the influence of educational programs across various measures in different age groups. The current study corroborates that bilingual educational programs vary in how much Spanish and English is used for instruction, even within nominal program type (Cirino et al., 2007; Saunders et al., 2006). These differences go beyond simple mean differences in the classroom-level outcomes: bilin-

gual programs may have different cross-language correlations, covariances, and variances (i.e., homogeneity of variance may be untenable). The current approach illustrates that these program effects occur at the classroom level and may imply complex differences. Interestingly, the multilevel approach allows us to examine the student-level relations after removing these classroom-level differences. The current results suggest that student relations are remarkably similar across programs after removing classroom-level differences due to instructional language and by respective educational program. Thus, it is possible that the grade trends in Figures 1–2 are a result of program or instructional effects, which may be clarified in future studies.

With regard to the use of instructional language, teachers, on average, adhered to their program labels, with immersion having predominantly English instruction. However, some immersion classrooms spent more than half of their instruction time in Spanish. The pattern of relations to classroom vocabulary performance was as expected: more time in a language contributed positively to that language's vocabulary and vice versa. Although there were too few classrooms to have much power, the estimates suggest substantial effects. The balance of English versus Spanish language used during instruction was found to explain up to 53% of the variability in classroom-level vocabulary outcomes. This further implies that instruction can change the cross-language relations between Spanish and English vocabulary. It is important to note that when instruction occurs for the classroom as a group, then those instructional effects need to be located at their appropriate level.

In general, the percentage of instructional English appeared to have larger effects for picture vocabulary than for narrative measures, suggesting that discrete word knowledge may be more easily influenced than narrative word use. Instructional language may therefore change not only mean classroom performance but also the relative effects of some tasks on other tasks at the classroom level. Such differences may have important implications for hypotheses of cross-language transfer (Bialystok, 2007; Bialystok et al., 2005; Carlo, 2001; Carlo & Royer, 1993; Durgunoglu, Nagy, & Hancin-Bhatt, 1993; Lindsey et al., 2003; Manis, Lindsey, & Bailey, 2004), as it is suggested here that such effects are at least partially due to classrooms rather than to students.

Measurement

The third major finding was that vocabulary measures differ widely and appear to be the strongest determinant of cross-language vocabulary correlation. In the prior literature shown in Figure 1, the level of Spanish–English correlation appears most strongly affected by which type of measure was used. Whereas both the narrative measures and WLPB vocabulary tests are expressive, the WLPB picture vocabulary test has scoring requirements that involve measuring knowledge of specific words rather than broad-based word knowledge, which might also include other verbal abilities (Peña, 2007; Peña & Kester, 2004). In this way, the WLPB could be said to be more focused on specific word knowledge, whereas the receptive PPVT may measure more general familiarity with words as well as more broad verbal abilities. Narrative measures, depending on their scoring, may measure even broader verbal abilities, given their positive relations across language. The correlations suggest this, in that the WLPB correlations tended to be negative (−.47 to .15), the correlations for the PPVT were more balanced (−.23 to .23), and the correlations for narrative measures were moderately positive (.34 to .63).

In the current analysis, we separated the questions of convergent and discriminant validity to student and classroom levels. At the student level, after controlling for the balance of instructional language used, both picture and narrative vocabulary measures were moderately positively related within language. Across language, however, picture vocabulary was relatively unrelated across languages, whereas the narratives were moderately positively related. Although it could be said that this positive relation could indicate that vocabulary, as indexed via narratives, exhibits little discriminant validity, narratives may involve verbal abilities that are fairly language-general when examined in a multilevel model of this sort. Simply put, the narratives seem to be measuring a fundamentally different type of vocabulary construct.

The student-level distinction between the picture vocabulary and narrative measures implies that narrative tasks call on different abilities than do other expressive tasks (e.g., the WLPB), which, in turn, may also be different from receptive tasks (e.g., the PPVT/TVIP, not measured in the current study). This difference implies that either vocabulary knowledge itself is different in these response modes or the response modes also call upon additional, as yet unmeasured, abilities. Thus, it seems defensible to refer to expressive and receptive vocabulary as separate constructs (Henricksen, 1999; Nation, 2001; Ouellette, 2006; Wise et al., 2007).

It is important to remember that the current study used only measures of expressive vocabulary, the WLPB, and narrative elicitation. Investigation of convergent and discriminant validity would be greatly assisted by use of receptive measures, such as the PPVT/TVIP, and perhaps measures of listening comprehension.

At the classroom level, the within-language relations across measures were moderately positive in Spanish and highly positive in English, suggesting high consistency in vocabulary performance. Across language, the picture vocabulary measures were strongly negatively related, suggesting that classrooms were specialized in their performance. Whereas the instructional language variable explained much of the variance in some of the programs, differences remained due to instructional focus (e.g., targeted vocabulary instruction) as well as other compositional factors due to the classroom, school, and community. For example, the strong negative relation between Spanish and English picture vocabulary may indicate that classrooms specialize in their instruction and grouping of students: Classrooms that performed well in one language generally performed poorly in the other language. These results at the classroom level may also suggest differences in how classrooms are affected by, and make use of, school and community resources.

Limitations

The current study has some important limitations to consider. Only two methods of measuring vocabulary were used in the current study: WLPB and narrative number of different words. Use of more measures would enable more sophisticated examinations of method, language, and/or trait differences (e.g., Branum-Martin et al., 2006). In particular, it would be informative to extend the current approach to measures of receptive vocabulary to better address questions of discriminant validity with regard to the

receptive–expressive distinction. Moreover, aside from grade level, student-level effects were not specifically modeled in the current article. Multilevel investigations into the role of other student characteristics, such as linguistic proficiency (e.g., other oral language skills), economics, or home environment, might prove informative.

The current article presents a cross-sectional approach to differences in student and classroom performance and instruction. An important extension will be to examine these questions longitudinally, such as with a cross-classified multilevel model, where the fact that students change teachers over time can be accounted for. In such a framework, differential teacher, instructional, and program effects may better be addressed while simultaneously using students as their own longitudinal controls. The current study is also limited by unmodeled effects at the classroom level and higher, such as school, regional, and community differences.

Another limitation is that because of the sharing of some teachers across classrooms (e.g., English language development instructors as campus-wide resources), instruction was implemented here as a percentage of time, rather than as actual minutes. It might be instructive to know the relation of absolute time on task to classroom outcomes. In addition, instructional content, such as vocabulary, spelling, or comprehension, might have had some effects that were not considered here.

Implications

The current study has implications for research and practice. The implications for research are that the type of measure, classroom differences, bilingual program, and instruction can have profound influence and are important to keep at the forefront of inquiry into bilingual issues. Additionally, student-level effects need to be separated from classroom- and instructional-level effects, as these can be opposite in direction from each other. We hope that the identified issues in vocabulary performance and the multilevel methods can help clarify the way in which questions of bilingual ability and instruction are approached. For example, a claim that vocabulary knowledge is a "language-specific skill" involves issues of construct (e.g., expressive vs. receptive), measure (picture, narrative, or other), grade level, and instruction. Without clarification of such issues, claims about abilities, instruction, or cross-language transfer are difficult to evaluate. Claims about the effectiveness of bilingual education programs are similarly difficult to evaluate properly without the appropriate methodological and substantive detail discussed here. Overall, the strongest determining factor for correlations between Spanish and English vocabulary appears to be the choice of the measure used to indicate vocabulary knowledge. More research is needed to understand the nature of word knowledge by the use of multiple measures, especially in bilingual populations. The same holds true of other language constructs, such as phonology, listening comprehension, and reading.

The implications for practice are that despite program or curriculum labels (e.g., immersion or transitional), teachers use Spanish and English in differing amounts. In the current study, some transitional teachers were found to use as little Spanish as some of the English immersion teachers, and some English immersion teachers spent more than half of their time conversing in Spanish. More important, whereas the number of observed classrooms limited the statistical significance of the relations between language of instruction and classroom-level outcomes, the proportion of variance explained suggests that instruction matters for classroom-level outcomes. It is possible that specific word knowledge, as measured on picture vocabulary tests, may be more easily influenced by instructional language than on expressive narrative measures. Thus, it is possible that specific words, as measured on academic picture vocabulary tests, may need to be taught in the target language but that more general narrative strategies (such as finding alternative words or adding explanations) might generalize across languages. Overall, we know very little about the cross-language effects of vocabulary knowledge and instruction, but it is hoped that this work and other similar ventures will continue to broaden our understanding of the issues involved in education for bilingual students.

References

August, D., Carlo, M. S., Dressler, C., & Snow, C. E. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research & Practice, 20*(1), 50–57.

August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority children: A research agenda.* Washington, DC: National Academy Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* New York: Oxford University Press.

Baker, C. (2001). *Foundations of bilingual education and bilingualism* (3rd ed.). Clevendon, England: Multilingual Matters.

Białystok, E. (2007). Acquisition of literacy in bilingual children: A framework for research. *Language Learning, 57*(Suppl. 1), 45–77.

Bialystok, E., Luk, G., & Kwan, E. (2005). Bilingualism, biliteracy, and learning to read: Interactions among languages and writing systems. *Scientific Studies of Reading, 9*(1), 43–61.

Branum-Martin, L., Mehta, P. D., Fletcher, J. M., Carlson, C. D., Ortiz, A., Carlo, M. S., et al. (2006). Bilingual phonological awareness: Multilevel construct validation among Spanish-speaking kindergarteners in transitional bilingual education classrooms. *Journal of Educational Psychology, 98*(1), 170–181.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Carlisle, J. F., Beeman, M. M., Davis, L. H., & Spharim, G. (1999). Relationship of metalinguistic capabilities and reading achievement for children who are becoming bilingual. *Applied Psycholinguistics, 20,* 459–478.

Carlo, M. S. (2001). *Do reading skills transfer across languages? Examining the literature from a component process perspective on reading.* Washington, D C: U.S. Department of Education, Office of Bilingual Education and Minority Language Affairs.

Carlo, M. S., August, D., McLaughlin, B., Snow, C. E., Dressler, C., Lippman, D. N., et al. (2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39,* 188–215.

Carlo, M. S., & Royer, J. M. (1993). *Theoretical and methodological issues in the study of cross-language transfer of reading skills.* Philadelphia: University of Pennsylvania, National Center on Adult Literacy.

Carroll, J. B. (1983). Psychometric theory and language testing. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 80–107). Rowley, MA: Newbury House.

Carroll, J. B. (1986). Second language. In R. F. Dillon & R. J. Sternberg (Eds.), *Cognition and instruction* (pp. 83–125). San Diego, CA: Academic Press.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge, United Kingdom: Cambridge University Press.

Cirino, P. T., Pollard-Durodola, S. D., Foorman, B. M., Carlson, C. D., & Francis, D. J. (2007). Teacher characteristics, classroom instruction, and student literacy and language outcomes in bilingual kindergartners. *Elementary School Journal, 107,* 341–364.

Conboy, B. T., & Thal, D. J. (2006). Ties between the lexicon and grammar: Cross-sectional and longitudinal studies of bilingual toddlers. *Child Development, 77,* 712–735.

Cronbach, L. J. (1942). An analysis of techniques for diagnostic vocabulary testing. *Journal of Educational Research, 36,* 206–217.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49,* 222–251.

Cummins, J. (1983). Language proficiency and academic achievement. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 108–129). Rowley, MA: Newbury House.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research, 38,* 529–569.

Dickinson, D. K., McCabe, A., Clark-Chiarelli, N., & Wolf, A. (2004). Cross-language transfer of phonological awareness in low-income Spanish and English bilingual preschool children. *Applied Psycholinguistics, 25,* 323–347.

Dressler, C., & Kamil, M. (2006). First- and second-language literacy. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 197–238). Mahwah, NJ: Erlbaum.

du Toit, S. H. C., & du Toit, M. (2003). Multilevel structural equations models. In J. DeLeeuw & I. G. G. Kreft (Eds.), *Handbook of quantitative multilevel analysis* (pp. 273–321). Boston: Kluwer.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Dunn, L. M., Padilla, E. R., Lugo, D. E., & Dunn, L. M. (1986). *Test de Imágenes Peabody–Adaptación Hispanoamericana.* Circle Pines, MN: American Guidance Service.

Durgunoglu, A. Y., Nagy, W. E., & Hancin-Bhatt, B. J. (1993). Cross-language transfer of phonological awareness. *Journal of Educational Psychology, 85,* 453–465.

Eid, M., Lischetzke, T., & Nussbeck, F. W. (2006). Structural equation models for multitrait–multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 283–299). Washington DC: American Psychological Association.

Fernández, M. C., Pearson, B. Z., Umbel, V. C., Oller, D. K., & Molinet-Molina, M. (1992). Bilingual receptive vocabulary in Hispanic preschool children. *Hispanic Journal of Behavioral Sciences, 14,* 268–276.

Fiestas, C. E., & Peña, E. D. (2004). Narrative discourse in bilingual children: Language and task effects. *Language, Speech, and Hearing Services in Schools, 35,* 155–168.

Fitzgerald, J. (1995). English-as-a-second-language learners' cognitive reading processes: A review of research in the United States. *Review of Educational Research, 65,* 145–190.

Foorman, B. R., Goldenberg, C., Carlson, C. D., Saunders, W., & Pollard-Durodola, S. D. (2004). How teachers allocate time during literacy instruction in primary-grade English language learner classrooms. In P. McCardle & V. Chhabra (Eds.), *The voice of evidence in reading research* (pp. 289–328). Baltimore, MD: Brookes.

Francis, D. J., Carlson, C. D., Fletcher, J. M., Foorman, B. R., Goldenberg, C., & Vaughn, S. (2005). Oracy/literacy development of Spanish-speaking children: A multi-level program of research on language minority children and the instruction, school and community contexts, and interventions that influence their academic outcomes. *Perspectives, 31*(2), 8–12.

Genesee, F., & Geva, E. (2006). Cross-linguistic relationships in working memory, phonological processing, and oral language. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 175–183). Mahwah, NJ: Erlbaum.

Genesee, F., Geva, E., Dressler, C., & Kamil, M. (2006). Synthesis: Cross-linguistic relationships. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 153–174). Mahwah, NJ: Erlbaum.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Arnold.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53,* 455–467.

Gottardo, A. (2002). The relationship between language and reading skill in bilingual Spanish–English speakers. *Topics in Language Disorders, 22,* 46–70.

Graves, M. F. (1987). The roles of instrucion in fostering vocabulary development. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 165–184). Hillsdale, NJ: Erlbaum.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Henricksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition, 21,* 303–317.

Johnson, J. (1989). Factors related to cross-language transfer and metaphor interpretation of bilingual children. *Applied Psycholinguistics, 10,* 157–177.

Kaplan, D., & Elliot, P. R. (1997). A didactic example of multilevel structural equation modeling applicable to the study of organizations. *Structural Equation Modeling, 4*(1), 1–24.

Lindsey, K. A., Manis, F. R., & Bailey, C. E. (2003). Prediction of first-grade reading in Spanish-speaking English-language learners. *Journal of Educational Psychology, 95,* 482–494.

Littell, R. D., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, NC: SAS Institute.

Manis, F. R., Lindsey, K. A., & Bailey, C. E. (2004). Development of reading in grades K–2 in Spanish-speaking English-language learners. *Learning Disabilities Research & Practice, 19,* 214–225.

McCabe, A., & Bliss, L. S. (2004–2005). Narratives from Spanish-speaking children with impaired and typical language development. *Imagination, Cognition and Personality, 24,* 331–346.

McLaughlin, B., August, D., & Snow, C. E. (2000). *Vocabulary knowledge and reading comprehension in English language learners: Final performance report.* (No. R306f60077–97). Washington, DC: Office of Educational Research and Improvement.

Mehta, P. D., Foorman, B. R., Branum-Martin, L., & Taylor, W. P. (2005). Literacy as a unidimensional multilevel construct: Validation, sources of influence, and implications in a longitudinal study in Grades 1–4. *Scientific Studies of Reading, 9*(2), 85–116.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations models. *Psychological Methods, 10,* 259–284.

Miccio, A., Tabors, P. O., Páez, M. M., Scheffner Hammer, C., & Wagstaff, D. A. (2005). Vocabulary development in Spanish-speaking Head Start children of Puerto Rican descent. In J. Cohen, K. McAlister, K. Rostad, & J. MacSwan (Eds.), *ISB4: Proceedings of the 4th International Symposium on Bilingualism* (pp. 1614–1617). Somerville, MA: Cascadilla Press.

Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., & Francis, D. J. (2006). Oral language and reading in bilingual children. *Learning Disabilities Research & Practice, 21*(1), 30–43.

Miller, J. F., & Iglesias, A. (2003–2004). *Systematic analysis of English and Spanish language transcripts.* Madison, WI: Waisman Center, University of Wisconsin—Madison.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338–354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22,* 376–398.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

Nagy, W. E., García, G. E., Durgunoglu, A. Y., & Hancin-Bhatt, B. J. (1993). Spanish–English bilingual students' use of cognates in English reading. *Journal of Reading Behavior, 25,* 241–259.

Nagy, W. E., & Scott, J. A. (2000). Vocabulary processes. In M. Kamil, P. Mosenthal, P. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 269–284). Mahwah, NJ: Erlbaum.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge, United Kingdom: Cambridge University Press.

Oller, D. K. (2005). The distributed characteristic in bilingual learning. In J. Cohen, K. McAlister, K. Rostad, & J. MacSwan (Eds.), *ISB4: Proceedings of the 4th International Symposium on Bilingualism* (pp. 1744–1749). Somerville, MA: Cascadilla Press.

Oller, D. K., & Eilers, R. E. (Eds.). (2002). *Language and literacy in bilingual children.* Clevedon, England: Multilingual Matters.

Ordóñez, C. L., Carlo, M. S., Snow, C. E., & McLaughlin, B. (2002). Depth and breadth of vocabulary in two languages: Which vocabulary skills transfer? *Journal of Educational Psychology, 94,* 719–728.

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology, 98,* 554–566.

Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development, 78,* 1255–1264.

Peña, E. D., & Kester, E. S. (2004). Semantic development in Spanish–English Bilinguals: Theory, assessment, and intervention. In B. A. Goldstein (Ed.), *Bilingual language development and disorders in Spanish–English speakers* (pp. 105–128). Baltimore, MD: Brookes.

Proctor, C. P., August, D., Carlo, M. S., & Snow, C. E. (2006). The intriguing role of Spanish language vocabulary knowledge in predicting English reading comprehension. *Journal of Educational Psychology, 98,* 159–169.

Pythian-Sence, C., & Wagner, R. K. (2007). Vocabulary acquisition: A primer. In R. K. Wagner, A. E. Muse, & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 1–14). New York: Guilford Press.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69,* 167–190.

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage.

Read, J. (2000). *Assessing vocabulary.* Cambridge, United Kingdom: Cambridge University Press.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18,* 1–32.

San Francisco, A. R., Carlo, M. S., August, D., & Snow, C. E. (2006). The role of language of instruction and vocabulary in the English phonological awareness of Spanish–English bilingual children. *Applied Psycholinguistics, 27,* 229–246.

San Francisco, A. R., Mo, E., Carlo, M., August, D., & Snow, C. (2006). The influences of language of literacy instruction and vocabulary on the spelling of Spanish–English bilinguals. *Reading and Writing, 19,* 627–642.

Saunders, W. M., Foorman, B. R., & Carlson, C. D. (2006). Is a separate block of time for oral English language development in programs for English learners needed? *Elementary School Journal, 107,* 181–198.

Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy research* (Vol. 1, pp. 97–110). New York: Guilford Press.

Sénéchal, M., Ouellette, G. P., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 173–182). New York: Guilford Press.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.

Snow, C. E., & Kim, Y.-S. (2007). Large problem spaces: The challenge of vocabulary for English language learners. In R. K. Wagner, A. E. Muse, & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 123–136). New York: Guilford Press.

Swanson, H. L., Sáez, L., Gerber, M., & Leafstedt, J. (2004). Literacy and cognitive functioning in bilingual and nonbilingual children at or not at risk for reading difficulties. *Journal of Educational Psychology, 96,* 3–18.

Tabors, P. O., Páez, M. M., & López, L. M. (2003). Dual language abilities of bilingual four-year olds: Initial findings from the Early Childhood Study of Language and Literacy Development of Spanish-speaking children. *NABE Journal of Research and Practice, 1,* 70–91.

Thomas, W. P. (1992). An analysis of the research methodology of the Ramírez study. *Bilingual Research Journal, 16,* 213–245.

Tilstra, J., & McMaster, K. (2007). Productivity, fluency, and grammaticality measures from narratives: Potential indicators of language proficiency? *Communication Disorders Quarterly, 29,* 43–53.

Uccelli, P., & Páez, M. M. (2007). Narrative and vocabulary development of bilingual children from kindergarten to first grade: Developmental changes and associations among English and Spanish skills. *Language, Speech, and Hearing Services in Schools, 38,* 225–236.

Uchikoshi, Y. (2005). Narrative development in bilingual kindergarteners: Can *Arthur* help? *Developmental Psychology, 41,* 464–478.

Uchikoshi, Y. (2006a). Early reading in bilingual kindergarteners: Can educational television help? *Scientific Studies of Reading, 10,* 89–120.

Uchikoshi, Y. (2006b). English vocabulary development in bilingual kindergarteners: What are the best predictors? *Bilingualism: Language and Cognition, 9*(1), 33–49.

Umbel, V. M., Pearson, B. Z., Fernández, M. C., & Oller, D. K. (1992). Measuring bilingual children's receptive vocabularies. *Child Development, 63,* 1012–1020.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., et al. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. *Developmental Psychology, 33,* 468–479.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition, 30*(1), 79–95.

Wise, J. C., Sevcik, R. A., Morris, R. D., Lovett, M. W., & Wolf, M. (2007). The relationship among receptive and expressive vocabulary, listening comprehension, pre-reading skills, word identification skills, and reading comprehension by children with reading disabilities. *Journal of Speech, Language & Hearing Research, 50,* 1093–1109.

Woodcock, R. W. (1991). *Woodcock Language Proficiency Battery—Revised.* Itasca, IL: Riverside.

Woodcock, R. W., & Muñoz-Sandoval, A. F. (1995). *Woodcock Language Proficiency Battery—Revised, Spanish Form.* Chicago: Riverside.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives.* Ann Arbor: University of Michigan Press.

Received June 4, 2008
Revision received January 8, 2009
Accepted February 19, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 912–925

0022-0663/09/$12.00 DOI: 10.1037/a0016647

Teacher–Child Interactions and Children's Achievement Trajectories Across Kindergarten and First Grade

Timothy W. Curby
George Mason University

Sara E. Rimm-Kaufman and Claire Cameron Ponitz
University of Virginia

This study examined the extent to which the quality of teacher–child interactions and children's achievement levels at kindergarten entry were associated with children's achievement trajectories. Rural students (n = 147) were enrolled in a longitudinal study from kindergarten through first grade. Growth trajectories (initial level and slope) were modeled with hierarchical linear modeling for 3 areas of achievement: word reading, phonological awareness, and mathematics. Cross-classified analyses examined the extent to which quality of teacher–child interactions and children's starting level predicted achievement growth rates over 2 years, and they also accounted for the changing nesting structure of the data. Results indicated that achievement at kindergarten entry predicted children's growth for all 3 outcomes. Further, first-grade teachers' strong emotional support related to greater growth in students' phonological awareness. Emotional and instructional support in first grade moderated the relationship between initial achievement and growth in word reading. Kindergarten classroom organization moderated the relationship between initial achievement and growth in mathematics. The implications of schooling for early growth trajectories are discussed.

Keywords: kindergarten, first grade, teacher–child interactions, achievement gap, cross-classified

Growth trajectories show that children who are behind early in school tend to have more difficulty catching up in later years (Jimerson, Egeland, & Teo, 1999; Juel, 1988; McClelland, Acock, & Morrison, 2006). Existing variation in achievement levels and growth trajectories is most problematic in relation to the so-called achievement gap that exists between children entering school with low versus high levels of initial performance. Findings from a national sample (Early Childhood Longitudinal Study, Kindergarten Class) point to striking variability in achievement levels upon school entrance (see http://nces.ed.gov/ECLS/ for more details about the data set). For example, some children not only can recognize letters and numbers upon the transition to kindergarten but can also sight-read words, do simple addition, and show comprehension of number sequences (West, Denton & Germino-Hausken, 2000). Other children enter school with very few, if any, of these skills. Broadly construed, the achievement gap points to large performance differences among children entering school; some children enter school with substantial academic skills and knowledge, whereas others do not.

Efforts to reduce the achievement gap between high- and low-performing students focus on schools, classrooms, and teachers. Children spend more time in classrooms than anyplace other than their homes (Hofferth & Sandberg, 2001; Rutter, Maughan, Mortimore, Ouston, & Smith, 1979). Further, the explicit statement of purpose of the No Child Left Behind Act (2002) is "to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education." Pursuant to this goal, the No Child Left Behind Act explicitly states the importance of "closing the achievement gap between high- and low-performing children," especially "between disadvantaged children and their more advantaged peers." As a result, classrooms are an important context to examine for correlates and predictors of children's achievement as well as solutions to the achievement gap problem (Seidman, Tseng, & Weisner, 2006).

What aspects of classrooms and teachers are likely to be most important in predicting achievement? Existing research has implicated the role of both classroom and teacher factors related to student success (Chatterji, 2006; Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008). In particular, such research points to the important role of classroom processes—interactions among teachers and children—as stronger predictors of child outcomes than distal factors, such as teacher education (Early et al., 2007). Despite a growing body of work on the role of classroom processes, there is virtually no research on the extent to which classroom processes contribute, year to year, toward children's achievement trajectories, especially in relation to raising the achievement levels of low-performing students (Cochran-Smith & Zeichner, 2005; Early et al., 2006; Mashburn et al., 2008). In the present study, we examined the quality of kindergarten and first-

Timothy W. Curby, Department of Psychology, George Mason University; Sara E. Rimm-Kaufman and Claire Cameron Ponitz, Center for Advanced Study of Teaching and Learning, University of Virginia.

Timothy W. Curby and Claire Cameron Ponitz thank the Institute of Education Sciences, U.S. Department of Education, for its fellowship support through the University of Virginia (R305B040049 and R305B060009, respectively). This work was also funded by National Science Foundation–Developmental and Learning Sciences Grant 0418469 to Sara E. Rimm-Kaufman. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education or the National Science Foundation. The work presented herein was conducted in partial fulfillment of Timothy W. Curby's dissertation at the University of Virginia.

Correspondence concerning this article should be addressed to Timothy W. Curby, George Mason University, Department of Psychology 3F5, 4400 University Drive, Fairfax, VA 22030-4444. E-mail: tcurby@gmu.edu

grade teachers' interactions with children as a set of classroom processes expected to contribute to students' developmental trajectories in early reading and mathematics skills. Additionally, we examined whether high-quality classroom experiences were more or less important for participants who entered kindergarten with different skill levels.

Theoretical Perspective

The theoretical basis for the present work draws from Bronfenbrenner's bioecological model (Bronfenbrenner & Morris, 1998, 2006). This model considers four sources of influence on children's development: process, person, context, and time. In this model, the "primary engines of development" are proximal processes (Bronfenbrenner & Morris, 1998, p. 996). These are increasingly complex, regularly occurring, reciprocal interactions between children and other people, objects, and ideas (Bronfenbrenner, 2005, p. 6). Proximal processes investigated herein refer to the reciprocal interactions between teachers and children; such interactions are hypothesized to be the primary mechanism by which children learn in classrooms.

The other three components constrain and influence the proximal processes that take place and, therefore, influence development indirectly. The *person* component of the model refers to unique attributes that children bring with them to the environment. The primary person characteristic investigated here was children's prior achievement. *Context* refers to environmental influences on proximal processes. This refers to the idea that different contexts (e.g., one classroom vs. another) influence the nature of the proximal processes taking place. Finally, *time* refers to the temporal dimension; children need to be exposed repeatedly to proximal processes over extended periods in order for them to have influence. This study followed participants from kindergarten to first grade; we observed the interactions to which children were exposed in these two different classrooms with multiple observations per year. Use of this framework to investigate teacher–child interactions may uncover the mechanisms through which teachers influence their students' development (Rutter & Maughan, 2002).

Contributors to Early Achievement

The central objective of schools is to promote academic achievement. However, there are large disparities in school readiness as a product of children's early experience and biologically based characteristics (Lee & Burkham, 2002; Rimm-Kaufman, Pianta, & Cox, 2001). Cognitive ability, attentional skills, and prior knowledge all contribute to achievement (e.g., Duncan et al., 2007). These characteristics are rooted in experiences children have prior to school entry and represent important resources or reflect specific risk processes (e.g., maternal education, poverty, poor health) in the child's life (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2005; Bradley, Corwyn, McAdoo, & Coll, 2001; Hebbeler, Spiker, Mallik, Scarborough, & Simeonsson, 2003). Competence at the start of kindergarten represents the crux of the achievement gap: Wide individual differences in children's early home learning experiences are largely responsible for later, widening differences in achievement manifesting during the school years (Bradley & Corwyn, 2002; Morrison, Bachman, & Connor, 2005). Many children come to school with limited levels of skills, abilities, and family social and economic resources; therefore, careful research is required about the extent to which schools and classrooms can compensate for disparate starting levels of achievement.

Our first goal in the present study was to examine the relation between initial achievement at kindergarten transition and subsequent achievement trajectories through the first grade. There is empirical support for three possible patterns of association. One possibility is that differences at the start of schooling widen over time (Chatterji, 2006; McClelland et al., 2006). This phenomenon, commonly called the "Matthew effect," has been robustly documented in literacy learning (Juel, 1988; McCoach, O'Connell, Reis, & Levitt, 2006, p. 15). A second alternative is that differences in achievement could remain stable after children begin schooling (McCoach et al., 2006). A third alternative is that children with low initial achievement begin closing the gap with their peers (Hamre & Pianta, 2005; Wright, Horn, & Sanders, 1997). These three possibilities rest, in part, on experiences in early schooling. Thus, it is important to consider the nature of children's classroom experiences in kindergarten and first grade.

Teacher–Child Interactions

What occurs in the classroom that can shift children's achievement trajectories in positive directions? We posited that when teachers interact with students in positive ways (whether individually, in small groups, or as a whole class), these interactions have the potential to provide children with support for their learning and may predict positive deflections in achievement trajectories (Pianta, La Paro, & Hamre, 2008). Such interactions can be classified into three domains of support: emotional, organizational, and instructional.

Emotional Support for Learning

Emotional support refers to the ways in which teachers foster positive classroom climate, minimize negative climate, attend sensitively to individual student needs, and emphasize student interests and autonomy (Birch & Ladd, 1998; Howes & Hamilton, 1992; Hyson, Copple, & Jones, 2006). Emotionally supportive classrooms have teachers who notice when students are struggling or need extra support, either academically or socially, and then respond to those needs appropriately (Hamre & Pianta, 2007). Further, emotionally supportive teachers adapt their plans as the lesson unfolds and support children's independence and expression of ideas (Battistich, Schaps, Watson, & Solomon, 1996; Bredekamp & Copple, 1997; Kern & Clemens, 2007).

In a classroom with high emotional support, students are able to take chances in their learning because of the safe environment created through the teacher's sensitive, responsive interactions (Hamre & Pianta, 2007). Perhaps for this reason, high-quality emotional support has been linked with higher achievement (Pianta, La Paro, Payne, Cox, & Bradley, 2002) and lower levels of problem behaviors (Mashburn et al., 2008). Nationally, elementary classrooms are often rated as providing medium-to-high levels of emotional support (Hamre, Pianta, Mashburn, & Downer, 2007).

Classroom Organization

Classroom organization refers to proactive management of the classroom that ensures productive use of time and materials and supports student attention and behavior. Teachers who offer high

levels of classroom organization prevent classroom behavior problems through the use of proactive strategies, optimize learning opportunities, minimize wasted time, and guide children's attention to learning objectives. When misbehavior does occur, teachers show high-quality classroom organization by efficiently and effectively reestablishing order and reengaging students (Emmer & Stough, 2001). Teachers also use orienting statements about the sequence of events in the classroom and establish regular, predictable routines (Bohn, Roehrig, & Pressley, 2004). Finally, organization also includes the quality of students' classroom activities, because varied and engaging instruction helps students learn (Stipek & Byler, 2004).

Strong organizational interactions among teachers and students have been linked to children's engagement (Rimm-Kaufman, Curby, Grimm, Nathanson, & Brock, 2009), which in turn relates to higher academic achievement (Brophy & Good, 1986; Ponitz, Rimm-Kaufman, Grimm, & Curby, 2009). Further, when teachers spend more time orienting and organizing their classrooms early in the school year, students spend less time engaging in unproductive transitions during the rest of the year (Cameron, Connor, & Morrison, 2005). Worth noting, on average, elementary classrooms in the United States are often found to have medium-to-high levels of classroom organization (Hamre et al., 2007).

Instructional Support for Learning

Instructional support involves teachers' encouragement of higher order thinking as students learn new concepts, provision of constructive and specific feedback, and stimulation of children's use of language. Teachers who offer high-quality instructional support make connections to the real world and focus less on having their students learn a set of facts and more on helping students learn the facts as part of larger themes and concepts (Battistich et al., 1996; Bredekamp & Copple, 1997; Committee on Learning Research and Educational Practice, National Research Council, 1999). Another component of high-quality instruction, quality of feedback, occurs when teachers focus on expanding learning and understanding instead of simply indicating whether an answer was correct (Franke, Kazemi, & Battey, 2007). Teachers who provide high levels of instructional support also use appropriate language modeling during frequent conversations with children.

Higher quality instructional support measured in these ways has been linked to higher scores on standardized tests of mathematics and reading achievement in prekindergarten (Curby et al., 2009; Mashburn et al., 2008) as well as teacher-reported achievement in kindergarten (Pianta et al., 2002) and first grade (Hamre & Pianta, 2005). Higher levels of instructional support have also been linked to more observed on-task behavior (Pianta et al., 2002). However, although relatively high levels of emotional support and classroom organization occur in American classrooms, the typical American classroom provides low levels of instructional support (Hamre et al., 2007).

Various approaches exist in the assessment of classroom characteristics, including teacher report (e.g., Wachs, Gurkas, & Kontos, 2004), student report (e.g., Brody, Dorsey, Forehand, & Armistead, 2002), and, most germane to the present study, observational methods (e.g., Burchinal et al., 2008; Mashburn et al., 2008; National Institute of Child Health and Human Development Early Child Care Research Network [NICHD ECCRN], 2002). Among observational methods, there are both high- and low-inference measures. For example, within the NICHD Study of Early Child Care and Youth Development, multiple methods were used to assess the positive climate of the classroom (more information is available at https://secc.rti.org/). One method used a Likert scale to rate the positive emotional climate over a 10-min period (high inference). The other method used the number of times the teacher displayed positive affect in another 10-min period (low inference). Although there are advantages to each method, high-inference methods have the advantage of being able to take multiple indicators into account at the same time (Pianta et al., 2008) and, in this way, they may be more representative of children's experience. Perhaps for these reasons, there has been a growing interest in the use of higher inference classroom observation measures (e.g., Howes et al., 2008; Mashburn et al., 2008; Pianta et al., 2002). Because of this, we have used a high-inference observational measure to assess the nature of teacher–child interactions in early childhood classrooms. To further strengthen our ability to tap into students' experience in the present study, we used repeated observations over the course of each school year to achieve a clear picture of what kindergarten and first-grade children were experiencing.

Potential Moderating Effects of Teacher–Child Interactions

Thus far, we have emphasized classrooms as potentially positive contributors to children's achievement in a main effect framework. However, evidence indicates that particular teacher practices are not equally effective for all children (Connor, Morrison, & Katch, 2004). The most extensive literature in this vein focuses on literacy learning (Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998). Connor, Morrison, and Underwood (2007) found that children's growth in letter–word reading depended both upon their starting level and the type of instruction they received. Children with lower skills benefited more in classrooms where instruction focused on the basic building blocks of reading.

Maternal education (Hamre & Pianta, 2005), cognitive ability (National Institute of Child Health and Human Development Early Child Care Research Network & Duncan, 2003), and instructional contexts (Downer, Rimm-Kaufman, & Pianta, 2007) have been explored as moderators of the relation between teacher–child interaction quality and student outcomes. Studies have examined whether teacher–child interactions serve a protective role for children with limited skills because the interactions compensate for early experiences not present in the child's home life (Luthar, Cicchetti, & Becker, 2000). Yet, to our knowledge, no researchers have examined the potential moderating role of interactional quality on starting level to predict children's achievement trajectories over multiple years. In the present study, we sought to replicate and extend this work by examining whether the quality of teacher–child interactions moderates the relation between prior achievement and children's academic growth over 2 years.

Research Questions

The present study examined whether quality in teacher–child interactions differentially predicted children's academic trajecto-

ries over kindergarten and first grade on the basis of children's initial achievement. Four research questions and associated hypotheses were posed. First, is there child-level variability in the level of achievement at the beginning of kindergarten and in the rates of achievement growth? On the basis of the literature documenting individual differences in the achievement of American children (West et al., 2000), we expected significant variability in both level and growth rate in the three areas of achievement (word reading, sound awareness, and mathematics).

Second, do children's starting achievement levels predict children's growth rates? We expected that children who had higher levels of achievement at the beginning of kindergarten would grow at faster rates (i.e., we expected a Matthew effect).

Third, does quality of kindergarten and first-grade teacher–child interactions predict children's growth rates? We hypothesized that interactions with high-quality emotional, organizational, and instructional support would predict improved growth rates. Also, we expected that teacher–child interactions would be more predictive of literacy skills than math skills, because literacy instruction predominates in the early years of school.

Fourth, does the quality of kindergarten and first-grade teacher–child interactions moderate the association between children's academic starting level and growth rate? We expected, in line with work on the compensatory role of early schooling for disadvantaged children, that high-quality interactions would be more beneficial for children entering kindergarten with low levels of achievement than for children entering kindergarten with high levels of achievement.

Method

Participants

Data for the present inquiry were collected as part of a larger study examining the relative contributions of classroom supports and children's self-regulation over the kindergarten–first-grade transition. Children were recruited from seven schools in four rural districts serving a large number of poor and working-class families. During kindergarten registrations and open houses, parents were invited to enroll their children in the study. After agreeing and signing consent forms, parents completed a questionnaire that provided sociodemographic information on the family and child. The parents of 333 kindergarteners consented, and 4–7 children were selected from each classroom to participate in the study. The resulting kindergarten sample contained 171 children. These children were followed through first grade.

At kindergarten entry, children's mean age was 5.4 years (range 4.7–6.3 years). There were 79 female and 92 male children, 143 Caucasians, 23 African Americans, and 5 of other ethnicity. Family income was reported within $15,000 ranges; the modal family income fell between $15,000 and $29,000 (39 families) but varied from less than $15,000 (19 families) to more than $100,000 (10 families). Most mothers and fathers (101 mothers, 92 fathers) had indicated high school as their highest level of education. The final sample included 147 children who had achievement data available over both years of the study. These 147 children were not statistically different from the original sample in gender, prekindergarten experience at age 4, maternal education (high school education vs. more than high school education), or family income (less than $30,000 per year vs. more than $30,000 per year).

The 147 child participants were enrolled in 36 kindergarten classrooms. Kindergarten teachers averaged 18.1 years of teaching experience (range 1–37 years). Most of the kindergarten teachers (n = 31) had full certification and licensure; all of them had at least a bachelor's degree, and 11 held a master's degree. Of the kindergarten teachers, 35 were Caucasian and 1 was Hispanic; all were female. The 147 children had 37 different first-grade teachers. The first-grade teachers averaged 14.3 years of teaching experience (range 2–30 years). Almost all of the first-grade teachers (n = 35) had full certification and licensure; all of them had at least a bachelor's degree, and 4 also held a master's degree. Of the first-grade teachers, 35 were Caucasian and 2 were African American; one teacher was male.

Procedure

Data were gathered from three sources: parents (see above), children, and classroom observations. At the start of kindergarten, trained research assistants went to each school and administered achievement tests to all child participants. Researchers returned in the spring of kindergarten and the fall and spring of first grade to readminister the alternating forms of the achievement tests.

Research assistants blind to the purpose of the study conducted classroom observations throughout children's kindergarten and first-grade years. Each year was divided into three observation windows (fall: October–December; winter: January–March; spring: March–May). Each teacher was observed at least once during each observation window for a total of three to five times. The most common schedule was to have two cycles of observation on 1 day in the fall, another two cycles of observation in the winter, and then a final cycle of observation in the spring. In total, this amounted to 300–375 min of observation for each teacher.

Measures

Teacher–child interaction quality. The Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2004, 2008) measures the quality of interactions that teachers have with children. Ten observable dimensions of quality are scored on a 7-point Likert scale from 1 (*low*) to 7 (*high*). These 10 dimensions comprise three superordinate domains: emotional support, classroom organization, and instructional support.

Emotional support includes four dimensions. Positive climate measures the degree to which there is evidence for positive relationships between teachers and students (e.g., smiling, laughter, a general tone of respect in the classroom). Negative climate (reversed for analysis) captures teachers' negative affect (e.g., irritability, harsh tone) and the use of punitive control (e.g., yelling, threats, physical punishment). Teacher sensitivity primarily measures teachers' awareness of individual student needs and response to those needs. Regard for student perspectives encompasses teachers' flexibility and student focus (e.g., following students' lead in a discussion), support for children's autonomy (e.g., providing a choice among activities), and encouragement of students' expression. Across these four dimensions, internal consistency was strong across each year of the study (kindergarten, α = .93; first grade, α = .87).

Classroom organization includes three dimensions. Behavior management primarily focuses on teachers' use of proactive mea-

sures to avoid behavior problems (e.g., clear expectations) and the efficient redirection of misbehavior when it does occur. Productivity captures the degree to which teachers maximize learning time by providing activities for students to engage in (including when they are finished with one activity) and having brief transitions between activities. Instructional learning formats captures the degree to which teachers effectively involve students in classroom activities and use a variety of instructional modalities to maximize student interest. High internal consistency was evident across these three dimensions (kindergarten, $\alpha = .87$; first grade, $\alpha = .86$).

Instructional support comprises three dimensions. Concept development indicates the degree to which teachers push for more conceptual instead of rote understanding of the material by providing opportunities for students to analyze, reason, create, and integrate knowledge. Quality of feedback encompasses teachers' ability to promote student understanding primarily through the use of scaffolding and feedback loops (i.e., teacher–child interactions when teachers and children are responding to each other). Language modeling measures how much teachers promote conversation in the class and elaborate on students' language with more advanced language. Across these three dimensions, there was high internal consistency each year (kindergarten, $\alpha = .94$; first grade, $\alpha = .93$).

To create variables for analyses, we computed average scores for each teacher in each domain (emotional support, classroom organization, and instructional support) from the multiple ratings obtained at each observation time point (Pianta, La Paro, & Hamre, 2004, 2008). Thus, each teacher had a score with possible range from 1 to 7 for each domain. This score represented the average teacher–child interaction quality within each domain provided in that classroom.

The framework we used to evaluate interrater reliability is based on the CLASS authors' recommendations. Observers went through an intensive training program that involved 2 days of training and culminated in a test of reliability. In order for raters to be deemed reliable, 80% of codes (both within and between dimensions) must be within one scale point of agreement for a gold standard rating across five 20-min video segments of elementary classrooms. All raters met or exceeded this level of reliability. We took two additional steps to ensure that reliability was maintained throughout the study. First, weekly discussions were held to discuss any coding difficulties. The focus of these meetings was to have coders agree on ratings for ambiguous coding situations and to link their rationale to the CLASS coding manual. Second, we conducted drift tests at three points during the study using a similar procedure that had been used to establish reliability during the CLASS training. Again, coders met or exceeded the 80% reliability standard.

The CLASS originated in the NICHD Study of Early Child Care and Youth Development (as the Classroom Observation System) and has been used in several large-scale studies to measure the quality of teacher–child interactions in thousands of classrooms (e.g., NCEDL Multi-State and SWEEP studies; see also http://www.fpg.unc.edu/~ncedl/). The three-factor model has been validated by comparing the fit of one-, two-, and three-factor models with confirmatory factor analysis (Hamre et al., 2007). Of these, the three-factor model provided the best fit. These domains have been differentially related to children's development across several studies, with emotional support being related to social outcomes, classroom organization being related to some social outcomes and some achievement outcomes, and instructional support being related most strongly to children's academic outcomes (e.g., Howes et al., 2008; Mashburn et al., 2008; Pianta et al., 2002).

Academic achievement. Growth in children's academic achievement was measured with the Woodcock–Johnson III Tests of Achievement (WJ III; Woodcock, McGrew, & Mather, 2001). Three subtests were individually administered in the fall and spring of kindergarten and first grade. Letter–Word Identification was used to measure children's word reading achievement. Sound Awareness was used as a measure of phonological awareness. The other subtest, Applied Problems, was used to measure mathematical skills. The WJ III was chosen to tap aspects of children's early reading and mathematics ability that undergo significant development in the early years of formal schooling. The selected subtests of the WJ III correspond to the skills that are targeted in kindergarten and first-grade instruction: letter–word identification, sound awareness skills, and the ability to solve mathematical problems. The WJ III offers two alternate forms that were used for fall and spring testing.

At each time point, a *W*-score was created for each outcome. *W*-scores are transformations of the Rasch ability scores (McGrew & Woodcock, 2001). These scores take into account children's age at the time of the assessment and are vertically equated so they can be compared over time. The *W*-score for each subtest is centered on a mean of 500, which is the expected score for a 10-year-old North American child (Mather & Jaffe, 2002).

Data Analysis

Descriptive statistics were run, then analyses were pursued that examined preliminary associations among children's achievement variables, classroom membership, and classroom outcomes. We then used HLM 6.0 software (Raudenbush, Bryk, Cheong, & Congdon, 2004) to model growth trajectories with achievement measures nested with children and specified cross-classified hierarchical models with children nested within classrooms.

Variability in starting points and growth rates. The first research question examined variability in children's achievement growth during kindergarten and first grade. To address this question, we constructed a linear growth model for each outcome (Letter–Word Identification, Sound Awareness, and Applied Problems). Each model was specified as a two-level model with four achievement time points nested within each child (see the Appendix). We verified the simple structure of the data to be linear growth (vs. quadratic) by comparing model fit statistics. Each baseline linear model included random effects for both an initial starting level and a linear slope in achievement for each child. Statistical significance of the intercept and slope indicated child-level variability, such that there were individual differences in children's initial values as well as individual differences in children's growth trajectories. Empirical Bayesian estimates of the intercept and slope for each child were obtained with the HLM software and were treated as new child-level variables. The slope coefficient for each achievement test was used as the outcome, and the initial starting level was used as a predictor in further analyses.

Child and classroom predictors of growth rates. The remaining research questions used child achievement initial levels and classroom interaction quality to predict achievement growth (the slope variable). Because children changed classroom membership from kindergarten to first grade, traditional hierarchical linear

models could not be used with classroom-level predictors. Thus, a series of cross-classified random effects models (Raudenbush & Bryk, 2002) were used. This modeling feature, known as HCM2, is available in the HLM 6.0 program (Raudenbush et al., 2004). HCM2 is able to account for this data structure by allowing children to be classified as having one kindergarten teacher and having a separate classification with a different first-grade teacher. A cross-classified model functions similarly to HLM, except that it models Level 1 nested within two different Level 2 structures simultaneously. However, there is an important limitation of HCM2; it cannot yet account for a three-level model (measurements within child within two different classrooms). Therefore, slopes from the previous HLM analyses were used as the outcomes, such that Level 1 included child characteristics and Level 2 included predictors and random effects for kindergarten and first grade. Despite the fact that the kindergarten and first-grade teachers are not fully crossed, the model can still accommodate these data (Raudenbush & Bryk). In fact, Connor et al. (2007) used crossed classified models with 86 students nested within 40 first-grade and 33 second-grade classrooms. Our sample showed no model convergence problems that tend to occur when the sample size is too small.

The first cross-classified model specified was the unconditional model. This unconditional model is represented by the following equations:

$$\text{Level 1: } Y_{ijk} = \pi_{0jk} + e_{ijk} \quad (1)$$

$$\text{Level 2: } \pi_{0jk} = \theta_0 + b_{00j} + c_{00k} \quad (2)$$

Equation 1 states that the expected slope of child i who is in kindergarten classroom j and first-grade classroom k (Y_{ijk}) is equal to an average slope (π_{0jk}) plus an individual error associated with that particular child (e_{ijk}). Equation 2 states that the Level 1 average slope (π_{0jk}) is equal to a grand mean slope (θ_0) plus a random effect for kindergarten classroom j (b_{00j}) plus a random effect for first-grade classroom k (c_{00k}).

After these unconditional models were created, parsimonious final models were constructed for each achievement outcome, consistent with recommended practice for saving degrees of freedom in these complex models (Raudenbush & Bryk, 2002). In parsimonious model specification, a model is built in several iterations. Each predictor is added to the model separately. If a predictor is significant, it stays in the model. If it is not significant, it is removed. Thus, our final models include only those predictors that were significant. Main effects were entered first (e.g., kindergarten emotional support). After all main effects had been tested, each interaction was tested (e.g., Kindergarten Emotional Support × Initial Achievement). In other words, the final model included only significant predictors, which varied depending on the outcome. As a predictor of the slope, initial achievement was entered at Level 1. Other child factors (gender, maternal education) were entered but were never significant with initial achievement in the model. Then, main effects for classroom quality (emotional support, classroom organization, and instructional support) were entered separately for both kindergarten and first grade at Level 2. Finally, we examined whether kindergarten and first-grade classroom quality interacted with initial achievement to predict the slope. Next are the equations for the final model of the Applied Problems, in which kindergarten classroom organization moderates the relation between initial Applied Problems scores and the slope. Final cross-classified models for Letter–Word Identification and Sound Awareness are provided in the Appendix.

$$\text{Level 1: } Y_{ijk} = \pi_{0jk} + \pi_1 \text{ (initial)} + e_{ijk} \quad (3)$$

$$\text{Level 2: } \pi_{0jk} = \theta_0 + b_{00j} + c_{00k} \quad (4)$$

and

$$\pi_1 = \theta_1 + \gamma_{11} \text{ (K organization)} \quad (5)$$

The main distinction from the unconditional model equations is in the addition of Equation 5. In this equation, the estimate for the contribution of initial achievement to the slope, π_1, is shown to have a main effect and a moderated effect. As Equation 5 shows, there is a main effect for initial achievement, θ_1, and the effect of initial achievement varies as a function of kindergarten classroom organization, γ_{11}.

Results

A preliminary analysis of variance revealed significant differences in initial achievement according to kindergarten classroom membership for Letter–Word Identification, $F(35, 111) = 1.64$, $p < .03$; Sound Awareness, $F(35, 111) = 1.68$, $p = .02$; and Applied Problems, $F(35, 111) = 3.00$, $p < .001$. These differences suggested some mechanism of tracking children into classrooms. However, when children's achievement starting level was correlated with aspects of kindergarten teachers' quality, only two statistically significant and positive correlations emerged (out of a possible nine correlations). These were between the starting levels of Letter–Word Identification ($r = .19$, $p < .02$) and Sound Awareness ($r = .20$, $p < .02$) with kindergarten classroom organization.

Descriptive statistics for the predictors and outcomes are presented in Table 1. The average estimated initial scores were 358.36 for Letter–Word Identification, 457.70 for Sound Awareness, and 425.69 for Applied Problems, and there was an average monthly growth rate of 5.21, 1.69, and 1.90 points per month, respectively. On the CLASS 7-point scale (with 1, 2 = *low*; 3, 4, 5 = *moderate*; and 6, 7 = *high quality*), kindergarten and first-grade classrooms typically provided moderate levels of emotional support and classroom organization but low levels of instructional support. One noticeable difference was that instructional support was significantly higher in kindergarten ($M = 3.00$) than in first grade ($M = 2.36$), $t(71) = 3.80$, $p < .001$). This

Table 1
Descriptive Statistics for Predictors and Outcomes

Statistic	*M*	*SD*	Range
Letter–Word Identification, initial	358.36	16.38	321.33–418.49
Letter–Word Identification, slope	5.21	0.28	4.67–5.91
Sound Awareness, initial	457.70	9.74	435.38–479.93
Sound Awareness, slope	1.69	0.14	1.23–2.10
Applied Problems, initial	425.69	12.88	386.79–456.77
Applied Problems, slope	1.90	0.19	1.47–2.61
Kindergarten			
Emotional support	4.77	0.94	1.92–6.25
Classroom organization	4.23	1.03	1.97–5.94
Instructional support	3.00	0.84	1.60–5.13
First grade			
Emotional support	4.94	0.56	3.54–5.96
Classroom organization	4.27	0.79	2.70–5.81
Instructional support	2.36	0.58	1.60–4.59

indicates that kindergarten teachers' emotional support and classroom organization were not statistically different from those of first-grade teachers but that kindergarten teachers had higher levels of instructional support. Additionally, the standard deviations tended to be smaller across first grade than kindergarten, although these differences were statistically significant only for emotional support, Levene's $F(1, 71) = 8.16$, $p < .01$. This result indicates there was more variability in the quality of teachers' emotionally supportive interactions during kindergarten than first grade.

Child-Level Variability in Achievement Growth

Growth curves were first modeled for each achievement outcome, with the four measurements of achievement nested within each child. Results are presented for all three outcomes in Table 2. Random intercept and slope coefficients were significant for all three outcomes, and this indicated that children varied in their starting points and their growth rates over kindergarten and first grade. On the basis of these models, for Letter–Word Identification, children entered kindergarten scoring on average 358.36 *W*-score points and grew at an average rate of 5.21 *W*-score points per month. Likewise, for Sound Awareness, children's mean initial score was 457.70 *W*-score points, and children grew at 1.69 *W*-score points per month. For Applied Problems, children, on average, scored 425.69 *W*-score points at kindergarten entry and grew at a rate of 1.90 *W*-score points per month.

The variability in children's initial achievement and growth was rather large from a practical standpoint (see Table 1). For example, for Letter–Word Identification, the standard deviation was 16.38 *W*-score points. Given that the average child improved by 5.21 *W*-score points per month, two children whose scores differed by one standard deviation were approximately 3 months apart in skill level. Along the same lines, children one standard deviation apart on Sound Awareness and Applied Problems were over 5 months and 6 months apart in skills, respectively.

Children's Starting Level Predicting Growth Rate

The slope values from the growth curves for each child were used as the outcomes in all subsequent analyses. For all three outcomes, children's starting achievement levels were significant predictors of individual growth rates. In other words, the skills with which children entered kindergarten predicted how quickly children improved. Surprisingly, the associations were not all in the same direction. For Letter–Word Identification, starting level was positively associated with children's growth, so that children at a higher starting level showed more growth; this was consistent with our hypothesis. For Sound Awareness and Applied Problems, however, the relation was in the opposite direction. The coefficient for the initial score was statistically significant and negative for both Sound Awareness (−.007) and Applied Problems (−.01). This indicated that children who started higher grew at a slightly slower rate.

The magnitude of these associations can be assessed relative to the growth rates. If we use Letter–Word Identification as an example, by multiplying the standard deviation (16.38) by the initial score coefficient (.01), we can determine that a child one standard deviation below the mean grew at a rate of 0.16 *W*-score units per month slower, relative to the growth rate for children with mean starting values (5.21). In other words, the modeled difference between two children starting 16.38 points apart in kindergarten was 19.26 points apart by the end of first grade. Based solely on differences in growth due to initial scores, the child starting behind initially would be approximately 2 weeks further behind at the end of first grade.

The estimates for Sound Awareness and Applied Problems outcomes were in the other direction. If we use the same method, for Sound Awareness, children starting one standard deviation below the mean gained on the average-scoring child by almost 3 weeks. But it is important to keep in mind that, given initial differences, these students would still be about 5 months apart at the end of first grade. For Applied Problems, children starting one standard deviation below the mean gained on the average-scoring child by about a week but were still over 6 months behind at the end of first grade.

Teacher–Child Interactions and Children's Growth Rate

To assess the third research question, all models tested for main effects of kindergarten and first-grade teacher–child interactions on children's growth rates. However, only one significant main effect for teacher–child interactions emerged. There was a positive relation between first-grade emotional support and Sound Awareness growth (see Table 3). For every 1-point increase in emotional support, children's rate of growth increased by 2.4% (coefficient for emotional support [.04] divided by the growth rate [1.69]). It is worth estimating the typical cumulative gains associated with a 1-point increase in

Table 2
Children's Achievement Growth Over Kindergarten and First Grade

	Letter–Word Identification			Sound Awareness			Applied Problems		
Statistic	Coefficient	*df*	*t* ratio	Coefficient	*df*	*t* ratio	Coefficient	*df*	*t* ratio
Fixed effect									
Intercept	358.36	146	195.31***	457.70	146	466.51***	425.69	146	328.52***
Slope	5.21	146	50.92***	1.69	146	36.73***	1.90	146	31.44***
	Variance	*df*	χ^2	Variance	*df*	χ^2	Variance	*df*	χ^2
Random effects									
Intercept	326.85	146	432.89***	110.03	146	656.94***	192.34	146	663.90***
Slope	0.28	146	176.49*	0.08	146	190.14**	0.13	146	189.79**
Level 1 effects, *r*	226.52			41.89			73.32		

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 3
Predictors of Children's Sound Awareness Growth

Predictor	Sound Awareness slope		
	Coefficient	*df*	*t* ratio
Fixed effects			
Average slope (intercept)	1.69	144	157.92***
Initial score on test	−0.007	144	−5.99*
First-grade emotional support	0.04	144	2.17*
	Variance	*df*	χ^2
Random effects			
Kindergarten (row)	<.001	34	42.62
First grade (column)	<.001	35	26.51
Level 1	.017		

Note. The intercept in this cross-classified model is actually a slope representing gains in *W*-score points per month.
* $p < .05$. *** $p < .001$.

emotional support over the course of a school year; an additional 0.04 *W*-score units per month equals a difference of 0.32 *W*-score units at the end of the academic year. Consideration of this 0.32 difference in light of the growth rate of 1.69 indicates that every 1-point difference of emotional support was associated with about a 1-week difference in phonological skills by the end of the year.

Moderating Effects of Teacher–Child Interactions on Children's Starting Level and Growth Rate

Teacher–child interactions moderated the relation between initial achievement and growth both for Letter–Word Identification and for Applied Problems. For Letter–Word Identification, two significant interactions were identified (see Table 4). First-grade instructional support and first-grade emotional support each interacted with children's initial achievement level. Figure 1 shows that

Table 4
Predictors of Children's Letter–Word Identification Growth

Predictor	Letter–Word Identification slope		
	Coefficient	*df*	*t* ratio
Fixed effects			
Average slope (intercept)	5.22	143	270.68***
Initial score on test	0.01	143	10.79***
Initial Score × First-Grade Emotional Support	0.006	143	2.95**
Initial Score × First-Grade Instructional Support	−0.005	143	−2.01*
	Variance	*df*	χ^2
Random effects			
Kindergarten (row)	<.001	35	38.78
First grade (column)	.002	36	46.12
Level 1	.038		

Note. The intercept in this cross-classified model is actually a slope representing gains in *W*-score points per month.
* $p < .05$. ** $p < .01$. *** $p < .001$.

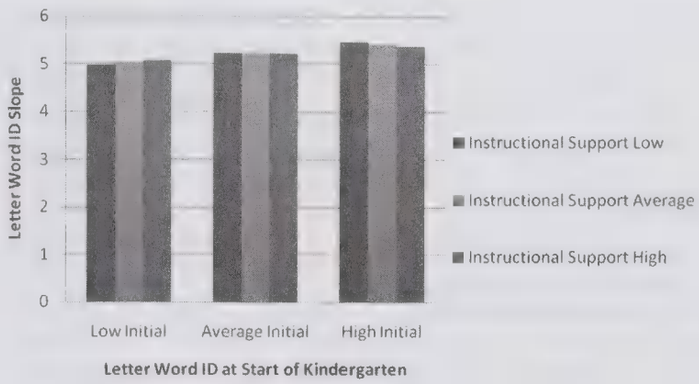


Figure 1. Moderating effects of first-grade instructional support. Low = −1 *SD*; High = +1 *SD*.

for first-grade instructional support, children with lower initial Letter–Word Identification scores benefited the most from higher quality instructional support but that children with higher initial scores benefited the most from lower quality instructional support. Figure 2 shows that for first-grade emotional support, children with higher initial Letter–Word Identification scores scored higher in classrooms with higher quality emotional support but that children with lower initial scores scored higher in classrooms with lower quality emotional support.

Table 5 presents results for Applied Problems. There was one significant interaction between the initial score and kindergarten classroom organization (see Figure 3). Children with lower initial Applied Problems scores grew at a faster rate in classrooms with higher quality classroom organization. Conversely, children with higher initial scores grew the fastest in classrooms with lower quality classroom organization.

The magnitude of effects on children's growth rates for all statistically significant interactions was small. In general, the effect sizes of the interactions are eclipsed by the differences between the low versus high initial ability (±1 *SD*). For example, for Letter–Word Identification, the difference between these higher and lower scoring groups was about 33 *W*-score points and the difference between the growth rates between children receiving high or low instructional quality (±1 *SD*) was about 0.1 *W*-score points per month. Thus, the accumulated benefit over an entire

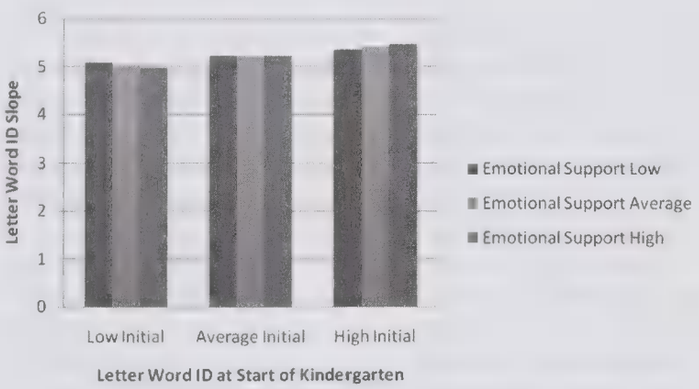


Figure 2. Moderating effects of first-grade emotional support. Low = −1 *SD*; High = +1 *SD*.

Table 5
Predictors of Children's Applied Problems Growth

Predictor	Applied Problems slope		
	Coefficient	*df*	*t* ratio
Fixed effects			
Average slope (intercept)	1.90	144	141.86***
Initial score on test	−0.01	144	−11.87***
Initial Score × Kindergarten Classroom Organization	−0.003	144	−2.89**
	Variance	*df*	χ^2
Random effects			
Kindergarten (row)	<.001	35	43.78
First grade (column)	.002	36	49.58
Level 1	.016		

Note. The intercept in this cross-classified model is actually a slope representing gains in *W*-score points per month.
** $p < .01$. *** $p < .001$.

school year (8 months) was an additional gain of less than 1 *W*-score point.

Discussion

We examined the extent to which classroom processes contributed to children's trajectories in three different achievement outcomes over kindergarten and first-grade year. Four findings emerged: (a) Children's initial levels and growth rate varied significantly in word reading, phonological awareness, and mathematics over kindergarten and first grade; (b) achievement at kindergarten entry strongly predicted children's rate of growth; (c) first-grade teachers' emotional supportiveness contributed to growth in phonological awareness; and (d) the contribution of interaction quality to the development of word reading and mathematics depended partially on children's prior skill level.

Variability in Children's Initial Achievement and Growth

For all three achievement outcomes tested, children varied in starting level of achievement and in their rate of growth during the kindergarten and first-grade years. There were large, practically important differences in children's initial achievement; a child initially scoring below average (−1 *SD*) appeared 3–6 months behind, compared to a child initially scoring at the average. Also, children exhibited notable disparities in their growth trajectories over the first 2 years of school. If differences in initial achievement were controlled for, differences in growth rates alone could lead to children being several weeks apart in their learning by the end of first grade. Projecting forward, these differences in growth could be quite consequential for children's later schooling.

The initial differences among children, combined with their different rates of growth, mean that kindergarten and first-grade teachers are faced with the challenge of tailoring instruction to the learning needs of a number of students who may be a year or more apart in their academic skills. The diversity in initial levels of achievement combined with variability in growth during the early years speaks to the challenge of meeting some of the primary objectives established in the No Child Left Behind Act, including closing the achievement gap between disadvantaged children and their privileged counterparts and providing all students with high-quality instruction.

Achievement at Kindergarten Entry

Children's achievement at kindergarten entry predicted children's growth rates, although the nature of the relation (positive or negative) varied by domain. Consistent with the Matthew effect hypothesis, for word reading, participants who started higher grew faster, whereas those who started lower grew more slowly. That is, children's scores became more divergent those of other children over time. In contrast, for phonological awareness and mathematics, students who started higher grew more slowly and those who started lower grew faster. In other words, for these achievement domains, children's scores converged somewhat over time.

These varying associations are likely due to the skills themselves and to classroom affordances for developing those skills. Letter–Word Identification tests children's early word reading abilities, which are a central focus in the early years of school. Because of the many reading opportunities in a classroom, children who enter school with some reading proficiency may be better able to improve their skills than are those who cannot read. In contrast, Sound Awareness tests phonological knowledge, a foundational set of skills that underlie reading. Teachers may focus more of their efforts on children who lack basic phonological skills, whereas average- and higher achieving children may not receive as much attention. Similar processes could be operating that would explain our results for Applied Problems, a measure of mathematical ability. For example, a teacher may spend more time with children who do not know their numbers than with students who are proficient in this area. As a result, higher achieving children may not improve as quickly.

The Contribution of Emotional Support to Early Phonological Skills

The quality of emotional support offered by first-grade teachers was associated with faster rates of growth phonological awareness, regardless of children's achievement at kindergarten entry. That is, higher levels of quality of first-grade emotional support were associated with higher rates of growth in Sound Awareness. Sur-

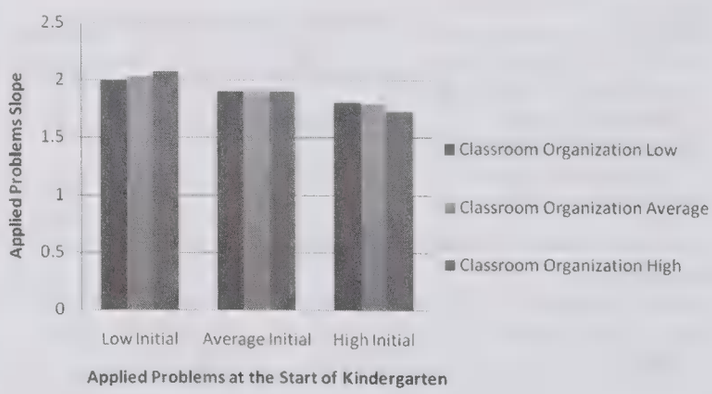


Figure 3. Moderating effects of kindergarten classroom organization. Low = −1 *SD*; High = +1 *SD*.

prisingly, this was the only main effect found for emotional support, classroom organization, or instructional support across all three achievement domains.

Associations between emotional support and children's achievement have been documented previously (Pianta et al., 2002, 2008). One explanation has to do with children's connectedness to school. Teachers who offer more emotional support are warm and responsive, and they foster relationships with students that would facilitate children's connectedness to school. Such relational closeness and school connectedness foster an environment conducive to learning (Baker, 2006; Birch & Ladd, 1998). An alternative explanation has to do with the phonological skills measured by the Sound Awareness subtest, which represents a complex set of related microskills shown to be receptive to direct instruction (Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001; Runge & Watkins, 2006). In particular, the Sound Awareness test taps a child's ability to substitute sounds for one another (e.g., substituting the /b/ sound in *bow* with /sl/ to make the word *slow*). To help students learn such specific skills, teachers must attend to individual children's knowledge of the sounds that comprise words. Teachers who are sensitive will notice that a child is having difficulty, determine the reason for the difficulty, and respond appropriately (Connor, Son, Hindman, & Morrison, 2005).

Effective Supports for Children's Development Depended on Their Initial Achievement

The present study indicates that development in the early years of school depends in part on children's initial skill level. Some findings were counterintuitive and were specific to some domains of classroom quality but not others. For word reading, children who entered kindergarten with lower scores had greater growth in classrooms with better first-grade instructional interactions but lower levels of emotional support. For mathematics, lower achieving children had greater growth in classrooms with better kindergarten classroom organization. The most likely explanation for this pattern is that the prediction of multiyear trajectories differs from prediction of within-year trajectories. Multiyear trajectories represent children's cumulative experience and thus may be more difficult to deflect.

Skills related to Letter–Word Identification were particularly sensitive to children's starting level (Connor, Morrison, & Petrella, 2004). Letter–Word Identification largely captures word reading abilities or orthographic knowledge (Rayner et al., 2001). In the present analyses, both first-grade emotional and instructional support were significant moderators of the relationship between children's initial levels of ability and growth. For word reading, students who had more to gain (those with low initial achievement) benefited the most from high-quality instruction. Surprisingly, participants with higher initial achievement benefited more from lower quality instruction. We suggest that, given that the mean of instructional support was so low, even in the higher quality classes, the small relative increase in concept development or language modeling may still have targeted the lower achieving students. In other words, relatively higher quality teachers in this sample were still not having high-quality instructional interactions (in absolute scores as measured by the CLASS). Thus, these counterintuitive findings could be a result that lacks generalizability. It is not clear whether the association between higher initial achievement level and lower quality instruction would prevail in a group of classrooms with higher levels of instructional support. This is an issue worthy of further inquiry.

Analyses also revealed that children with higher initial word reading scores showed greater gains in word identification in classrooms with higher levels of emotional support. In contrast, children with lower initial word reading showed greater gains in classrooms with lower levels of emotional support. The present findings are consistent with other work linking emotional support with student achievement. For example, Connor et al. (2005) found that students who had more warm and responsive teachers had better vocabulary and word-decoding skills. The present finding extends earlier work suggesting that emotional support may be even more important for children with higher, rather than lower, levels of initial ability. One explanation for this finding is that in classrooms with higher levels of emotional support, teachers tend to follow children's leads rather than adhere rigidly to their own plans and ignore student input. If the teacher is following the lead of high-achieving children, this may not benefit the lower achieving children. For example, if during a whole class activity a higher achieving child asks the teacher about the meaning of an advanced word that the lower achieving child does not recognize, students with lower ability will be unlikely to benefit from the answer.

In terms of mathematics achievement, children with lower initial ability showed more growth when they were in kindergarten classrooms with more classroom organization. In contrast, those with higher initial mathematics achievement grew more in classrooms with lower kindergarten classroom organization. This counterintuitive finding should be interpreted in context of the small percentage of instructional time typically devoted to mathematics during kindergarten (~11%; La Paro et al., in press). What does good classroom organization do for learning, and why might it be more important for some children than others? Better classroom organization may contribute to more mathematics learning time (Rimm-Kaufman et al., 2009) and in turn relate to gains in mathematics achievement for low math achievers. The association between classroom organization and mathematics learning time may be less consequential for children with high levels of mathematics achievement. Higher achieving students may actually function well in poorly managed classes. For example, a child who already knows math in a poorly organized classroom might choose to work independently on activities, whereas children with poor math skills might use the free time to disengage academically. It is also possible that high math achievers in a more organized class may receive more exposure to math but that the activities might still be too simple to challenge them.

Implications

Three primary implications follow from these findings. First, of the three interactions and one main effect for classroom processes, three were from first grade. Taken together, these findings suggest the importance of first grade, relative to kindergarten, for deflecting children's achievement trajectories (Perry, Donohue, & Weinstein, 2007). It has been found that developmental trajectories are established early in children's school careers (Juel, 1988) and become increasingly more stable as students progress through school (Alexander & Entwisle, 1988). A body of work emphasizes first grade as a critical developmental context with more stringent pressures and higher academic expectations than kindergarten (Alexander & Entwisle, 1988). For example, first graders are likely to be exposed to more activities relevant to early reading and

mathematics content learning and achievement than are kindergarteners (La Paro et al., in press; NICHD ECCRN, 2002). Although the present design does not allow for causal inferences, our findings suggest the importance of first grade and thereby imply that experiences children had in first grade were powerful enough to override a trajectory established in kindergarten.

The second implication regards young students' differential experiences of quality. It is of note that all interactions were disordinal interactions (Glass & Hopkins, 1996). In other words, lower achieving children did comparatively better in one classroom condition, and higher achieving children did comparatively better in a different classroom condition. This fact suggests that children with low initial achievement may have had different classroom needs than did children with high initial achievement. It could have been the case that all children benefited from teacher–child interactions but that those with lower initial scores benefited more than students with higher initial scores. This raises a question: What type of teacher is effective in boosting the achievement of high achievers? Existing research suggests that only the most effective teachers can boost learning in the best students (Sanders & Horn, 1998). These results suggest that many classrooms may lack sufficient affordances for high-achieving (or gifted) children and that instruction may be aimed at the lower or average-achieving children. Another consideration is that the amount of time spent teaching in a particular academic domain might have differential effects for high- and low-achieving students. In the CLASS measure, time and quality are confounded. Research is currently emerging that examines how quality and quantity may interact to explain children's development (Pianta, Belsky, et al., 2008).

Third, this study adds to an emerging body of work examining the lasting effects of early school experiences (e.g., Magnuson, Ruhm, & Waldfogel, 2007; Sanders & Rivers, 1996). Many studies have found within-year associations between children's development and instruction. The present study examined whether teacher–child interactions from 1 year altered the achievement trajectory for students based on 2 years. Arguably, a greater number of associations, or stronger associations, could have been found by looking within year. These longitudinal effects have been emphasized in the preschool literature, in particular. For example, Belsky et al. (2007) found that higher quality teacher–child interactions during preschool were associated with higher assessed vocabulary in fifth grade. The results from the present study are compatible. That is, learning trajectories can be altered by a kindergarten or first-grade experience, but the extent of that alteration is contingent upon achievement level when children begin school.

Limitations

Two limitations require mention. First, consistent with the theoretical underpinnings of the measure (Hamre & Pianta, 2007), the CLASS domains are positively correlated with one another (Hamre et al., 2007). This can present a problem of multicollinearity, particularly when testing interactions. This concern was partially mitigated by the fact that predictors were centered and parsimonious model building techniques were used (Raudenbush & Bryk, 2002). Second, growth rates herein did not account for summer learning loss, which has been shown to disproportionately affect children at-risk for school failure (Burkam, Ready, Lee, & LoGerfo, 2004). Although the data supported linear growth, this may not have been true for all students (McCoach et al., 2006). Linear growth rates for children who had summer losses would have appeared lower than for students without summer losses. A study powered by more participants and time points should examine the possibility of a nonlinear growth trajectory.

Conclusion

This study indicates that teachers face a significant and practical challenge in their classrooms as they strive to support the learning of students with varied abilities. Improving teachers' emotional support toward children is one step toward this goal; it appears to have positive implications for young students' achievement in sound awareness. However, large disparities exist in children's academic skill levels and growth, and the kind of teaching that is appropriate for lower achieving students may be somewhat different from the type of teaching that is well suited for higher achieving students. Thus, a nuanced understanding of classroom quality requires that we consider not only what constitutes good-quality instruction but also what facets of quality are more or less important depending on children's level of achievement upon school entry.

Many policymakers view the reduction of the achievement gap as a primary goal of schools. The present study found that the lower achieving students grew faster than higher achieving students in two important domains, phonological and mathematics skills. This finding suggests that the playing field is being leveled in these areas but that it may be happening at the expense of higher achieving students—at least in the relatively poor, rural districts that we studied. This study raises the question as to whether teachers are helping lower achieving students catch up to the higher achieving students or whether higher achieving students are not provided sufficient supports for reaching their potential.

The present study points to the need for more work to be done that follows the contribution of teaching to children's learning for more than a single year. Indeed, very little work has been done that examines teachers' accumulated effects on children. Children grow and change each year. Some of those changes reflect biologically based patterns of development, whereas others occur as a consequence of children's interactions with important adults. The fact that children move in and out of various classrooms provides us with a unique perspective from which to understand the contribution of individual teachers to student achievement.

References

Alexander, K. L., & Entwisle, D. R. (1988). Achievement in the first 2 years of school: Patterns and processes. *Monographs of the Society for Research in Child Development, 53*(2, Serial No. 218).

Baker, J. A. (2006). Contributions of teacher–child relationships to positive school adjustment during elementary school. *Journal of School Psychology, 44,* 211–229.

Battistich, V., Schaps, E., Watson, M., & Solomon, D. (1996). Early findings from an ongoing multisite demonstration trial. *Journal of Adolescent Research, 11,* 12–35.

Belsky, J., Vandell, D. L., Burchinal, M., Clarke-Stewart, K. A., McCartney, K., & Owen, M. T. (2007). Are there long-term effects of early child care? *Child Development, 78,* 681–701.

Birch, S. H., & Ladd, G. W. (1998). Children's interpersonal behaviors and the teacher–child relationship. *Developmental Psychology, 34,* 934–946.

Bohn, C. M., Roehrig, A. D., & Pressley, M. (2004). The first days of

school in the classrooms of two more effective and four less effective primary-grades teachers. *Elementary School Journal, 104,* 269–287.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2005, December). *How changes in entry requirements alter the teacher workforce and affect student achievement* (Working Paper 11844). Cambridge, MA: National Bureau of Economic Research.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53,* 371–399.

Bradley, R. H., Corwyn, R. F., McAdoo, H. P., & Coll, C. G. (2001). The home environments of children in the United States: Part I. Variations by age, ethnicity, and poverty status. *Child Development, 72,* 1844–1867.

Bredekamp, S., & Copple, C. (Eds.) (1997). *Developmentally appropriate practice in early childhood programs* (Rev. ed.). Washington, DC: National Association for the Education of Young Children.

Brody, G. H., Dorsey, S., Forehand, R., & Armistead, L. (2002). Unique and protective contributions of parenting and classroom processes to the adjustment of African American children living in single-parent families. *Child Development, 73,* 274–286.

Bronfenbrenner, U. (2005). *Making human beings human: Bioecological perspectives on human development.* Thousand Oaks, CA: Sage.

Bronfenbrenner, U., & Morris, P. (1998). The ecology of developmental process. In W. Damon (Series Ed.) & R. M. Lerner (Vol. Ed.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (5th ed., pp. 993–1028). New York: Wiley.

Bronfenbrenner, U., & Morris, P. (2006). The bioecological model of human development. In R. M. Lerner (Ed.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (6th ed., pp. 793–828). Hoboken, NJ: Wiley.

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Macmillan.

Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher–child interactions and instruction. *Applied Developmental Science, 12,* 140–153.

Burkam, D. T., Ready, D. D., Lee, V. E., & LoGerfo, L. F. (2004). Social-class differences in summer learning between kindergarten and first grade: Model specification and estimation. *Sociology of Education, 77,* 1–31.

Cameron, C. E., Connor, C. M., & Morrison, F. J. (2005). Effects of variation in teacher organization on classroom functioning. *Journal of School Psychology, 43,* 61–85.

Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) kindergarten to first grade sample. *Journal of Educational Psychology, 98,* 489–507.

Cochran-Smith, M., & Zeichner, K. M. (2005). *Studying teacher education: The report of the AERA panel on research and teacher education.* Mahwah, NJ: Erlbaum.

Committee on Learning Research and Educational Practice, National Research Council. (1999). In M. S. Donovan, J. D. Bransford, & J. W. Pellegrino (Eds.), *How people learn: Bridging research and practice.* Washington, DC: National Academy Press.

Connor, C. M., Morrison, F. J., & Katch, E. L. (2004). Beyond the reading wars: The effect of classroom instruction by child interactions on early reading. *Scientific Studies of Reading, 8,* 305–336.

Connor, C. M., Morrison, F. J., & Petrella, J. N. (2004). Effective reading comprehension instruction: Examining Child × Instruction interactions. *Journal of Educational Psychology, 96,* 682–698.

Connor, C. M., Morrison, F. J., & Underwood, P. S. (2007). A second chance at second grade: The independent and cumulative impact of first- and second-grade reading instruction and students' letter–word reading skill growth. *Scientific Studies of Reading, 11,* 199–233.

Connor, C. M., Son, S., Hindman, A., & Morrison, F. (2005). Teacher qualifications, classroom practices, family characteristics, and preschool experience: Complex effects on first graders' vocabulary and early reading outcomes. *Journal of School Psychology, 43,* 343–375.

Curby, T. W., LoCasale-Crouch, J., Konold, T. R., Pianta, R., Howes, C., Burchinal, M., et al. (2009). The relations of observed pre-k classroom quality profiles to children's academic achievement and social competence. *Early Education and Development, 20,* 346–372.

Downer, J. T., Rimm-Kaufman, S. E., & Pianta, R. C. (2007). How do classroom conditions and children's risk for school problems contribute to children's engagement in learning? *School Psychology Review, 36,* 413–432.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., et al. (2007). School readiness and later achievement. *Developmental Psychology, 43,* 1428–1446.

Early, D. M., Bryant, D. M., Pianta, R. C., Clifford, R. M., Burchinal, M. R., Ritchie, S., et al. (2006). Are teachers' education, major, and credentials related to classroom quality and children's academic gains in pre-kindergarten? *Early Childhood Research Quarterly, 21,* 174–195.

Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R. H., Bryant, D., et al. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development, 78,* 558–580.

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36,* 103–112.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90,* 37–55.

Franke, M. L., Kazemi, E., & Battey (2007). Mathematics teaching and classroom practice. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (2nd ed.). Charlotte, NC: Information Age Publishing.

Glass, G. V., & Hopkins, K. D. (Eds.) (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon.

Hamre, B., & Pianta, R. C. (2005). Can instructional and emotional support in the first grade classroom make a difference for children at risk for school failure? *Child Development, 76,* 949–967.

Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In R. C. Pianta, M. J. Cox, & K. Snow (Eds.), *School readiness and the transition to kindergarten* (pp. 49–84). Baltimore: Brookes.

Hamre, B. K., Pianta, R. C., Mashburn, A., & Downer, J. (2007). *Building and validating a theoretical model of classroom effects in over 4,000 early childhood and elementary classrooms.* Retrieved June 1, 2008, from the Foundation for Child Development Web site: http://www.fcd-us.org/resources/resources_show.htm?doc_id=507559

Hebbeler, K., Spiker, D., Mallik, S., Scarborough, A., & Simeonsson, R. (2003). *Demographic characteristics of children and families entering early intervention.* Menlo Park, CA: SRI International.

Hofferth, S. L., & Sandberg, J. F. (2001). How American children spend their time. *Journal of Marriage and Family, 63,* 295–308.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly, 23,* 27–50.

Howes, C., & Hamilton, C. E. (1992). Children's relationships with child care teachers: Stability and concordance with parental attachments. *Child Development, 63,* 867–878.

Hyson, M., Copple, C., & Jones, J. (2006). Early childhood development and education. In K. A. Renninger & I. E. Sigel (Eds.), *Handbook of child psychology: Vol. 4. Child psychology in practice* (6th ed., pp. 3–47). Hoboken, NJ: Wiley.

Jimerson, S., Egeland, B., & Teo, A. (1999). A longitudinal study of

achievement trajectories: Factors associated with change. *Journal of Educational Psychology, 91,* 116–126.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80,* 437–447.

Kern, L., & Clemens, N. H. (2007). Antecedent strategies to promote appropriate classroom behavior. *Psychology in the Schools, 44,* 65–75.

La Paro, K. M., Hamre, B. K., LoCasale-Crouch, J., Pianta, R. C., Bryant, D. M., Early, D. M., et al. (in press). Quality in kindergarten classrooms: Observational evidence for the need to increase children's learning opportunities in early education classrooms. *Early Education and Development.*

Lee, V. E., & Burkham, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school.* Washington, DC: Economic Policy Institute.

Luthar, S. S., Cicchetti, D., & Becker, B. (2000). The construct of resilience: A critical evaluation and guidelines for future work. *Child Development, 71,* 543–562.

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). The persistence of preschool effects: Do subsequent classroom experiences matter? *Early Childhood Research Quarterly, 22,* 18–38.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., et al. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language and social skills. *Child Development, 79,* 732–749.

Mather, N., & Jaffe, L. E. (2002). *Woodcock–Johnson III: Reports, recommendations, and strategies.* New York: Wiley.

McClelland, M. M., Acock, A. C., & Morrison, F. J. (2006). The impact of kindergarten learning-related skills on academic trajectories at the end of elementary school. *Early Childhood Research Quarterly, 21,* 471–490.

McCoach, D. B., O'Connell, A. A., Reis, S. M., & Levitt, H. A. (2006). Growing readers: A hierarchical linear model of children's reading growth during the first 2 years of school. *Journal of Educational Psychology, 98,* 14–28.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock–Johnson III technical manual.* Itasca, IL: Riverside Publishing.

Morrison, F. J., Bachman, H. J., & Connor, C. M. (2005). *Improving literacy in America: Guidelines from research.* New Haven, CT: Yale University Press.

National Institute of Child Health and Human Development Early Child Care Research Network. (2002). The relation of first grade classroom environment to structural classroom features, teacher, and student behaviors. *Elementary School Journal, 102,* 367–387.

National Institute of Child Health and Human Development Early Child Care Research Network & Duncan, G. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development, 74,* 1454–1475.

No Child Left Behind Act of 2001, 115 Stat. 1425 (2002). [Pub. L. No. 107–110]. Retrieved June 18, 2008, from http://www.ed.gov/policy/elsec/leg/esea02/pg1.html#sec1001

Perry, K. E., Donohue, K. M., & Weinstein, R. S. (2007). Teaching practices and the promotion of achievement and adjustment in first grade. *Journal of School Psychology, 45,* 269–292.

Pianta, R., Belsky, J., Vandergrift, N., Houts, R. M., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal, 45,* 365–397.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2004). *Classroom Assessment Scoring System—K-3.* Unpublished manuscript.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System—K-3.* Baltimore: Brookes Publishing.

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *Elementary School Journal, 102,* 225–238.

Ponitz, C. C., Rimm-Kaufman, S. E., Grimm, K. J., & Curby, T. W. (2009). How kindergarten classroom quality translates into reading achievement: The critical role of student engagement. *School Psychology Review, 38,* 102–120.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Raudenbush, S., Bryk, A., Cheong, Y. F., & Congdon, R. (2004). *HLM 6: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2,* 31–74.

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology, 45,* 958–972.

Rimm-Kaufman, S. E., Pianta, R. C., & Cox, M. J. (2000). Teachers' judgments of problems in the transition to kindergarten. *Early Childhood Research Quarterly, 15,* 147–166.

Runge, T. J., & Watkins, M. W. (2006). The structure of phonological awareness among kindergarten students. *School Psychology Review, 35,* 370–386.

Rutter, M., & Maughan, B. (2002). School effectiveness findings 1979–2002. *Journal of School Psychology, 40,* 451–475.

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, A. (1979). *Fifteen thousand hours: Secondary schools and their effects on children.* Cambridge, MA: Harvard University Press.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12,* 247–256.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Seidman, E., Tseng, V., & Weisner, T. S. (2006). Social setting theory and measurement. In *Report and resource guide 2004–2005.* New York: William T. Grant Foundation.

Stipek, D., & Byler, P. (2004). The Early Childhood Classroom Observation Measure. *Early Childhood Research Quarterly, 19,* 375–397.

Wachs, T. D., Gurkas, P., & Kontos, S. (2004). Predictors of preschool children's compliance behavior in early childhood classroom settings. *Applied Developmental Psychology, 25,* 439–457.

West, J., Denton, K., & Germino-Hausken, E. (2000). America's kindergarteners: Findings from the Early Childhood Longitudinal Study, kindergarten class of 1998–1999: Fall 1998. *Education Statistics Quarterly, 2,* 7–13.

Woodcock, T. A., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III Tests of Achievement.* Itasca, IL: Riverside Publishing.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11,* 57–67.

Appendix

Specifications for Hierarchical Linear Models and Cross-Classified Models

The first step in the analyses involved analyzing unconditional growth models for children's achievement (Letter–Word Identification, Sound Awareness, and Applied Problems). These models accounted for growth across four time points (fall and spring of kindergarten and first grade). These unconditional models can be represented by the following equations:

$$\text{Level 1 Model: } Y_{ij} = \beta_0 + \beta_1(\text{time}) + r$$

$$\text{Level 2 Model: } \beta_0 = \gamma_{00} + u_0$$

$$\beta_1 = \gamma_{10} + u_1$$

The Level 1 equation models achievement across the four time points. It states that achievement, Y, for child i at time point j is equal to the intercept for that achievement measure (in the fall of kindergarten), β_0, plus slope, β_1, plus individual error. The Level 2 equations model between-child variance. They state that the intercept, β_0, is equal to a grand mean for the intercept, γ_{00}, plus an individually varying (random) effect, u_0. Likewise, the slope, β_1, is equal to a grand mean for the slope, γ_{10}, plus an individually varying (random) effect, u_1. Implicit in this model is that individually varying intercepts and slopes were computed for each child.

The intercept (initial) and slope values from the above models were used in the cross-classified analyses. The final cross-classified model for Applied Problems is presented in the main text. Herein, final models are presented for Letter–Word Identification and Sound Awareness. First, the final model for Letter–Word Identification is represented by the following equations:

$$\text{Level 1: } Y_{ijk} = \pi_{0jk} + \pi_1(\text{initial}) + e_{ijk}$$

$$\text{Level 2: } \pi_{0jk} = \theta_0 + b_{00j} + c_{00k}$$

$$\pi_1 = \theta_1 + \gamma_{11}(\text{first-grade emotional support}) + \gamma_{12}(\text{first-grade instructional support})$$

The Level 1 equation states that the expected Letter–Word Identification slope (Y_{ijk}) of child i who is in kindergarten classroom j and first-grade classroom k is equal to an average slope (π_{0jk}), plus an effect for where that child started (i.e., initial), π_1, plus an individual error associated with that particular child (e_{ijk}). The Level 2 equations state that the Level 1 average slope (π_{0jk}) is equal to a grand mean slope (θ_0) plus a random effect for kindergarten classroom j (b_{00j}) plus a random effect for first-grade classroom k (c_{00k}). Additionally, the statistical effect that the initial score has on the slope varies as a function of the first-grade emotional support, γ_{11}, and the first-grade instructional support, γ_{12}, the child received.

Likewise, presented below are the final cross-classified equations for Sound Awareness.

$$\text{Level 1: } Y_{ijk} = \pi_{0jk} + \pi_1(\text{initial}) + e_{ijk}$$

$$\text{Level 2: } \pi_{0jk} = \theta_0 + b_{00j} + c_{00k}$$

$$\pi_1 = \theta_1$$

The Level 1 equation states that the expected Sound Awareness slope (Y_{ijk}) of a child i who is in kindergarten classroom j and first-grade classroom k is equal to an average slope (π_{0jk}), plus an effect for his or her initial score, π_1, plus an individual error associated with that particular child (e_{ijk}). The Level 2 equations state that the Level 1 average slope (π_{0jk}) is equal to a grand mean slope (θ_0), plus a main effect for first-grade emotional support, plus a random effect for kindergarten classroom j (b_{00j}), plus a random effect for first-grade classroom k (c_{00k}).

Received July 14, 2008
Revision received May 21, 2009
Accepted May 22, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 926–937

0022-0663/09/$12.00 DOI: 10.1037/a0016586

Longitudinal Impact of Two Universal Preventive Interventions in First Grade on Educational Outcomes in High School

Catherine P. Bradshaw and Jessika H. Zmuda
Center for the Prevention of Youth Violence and Bloomberg School of Public Health, Johns Hopkins University

Sheppard G. Kellam
Bloomberg School of Public Health, Johns Hopkins University

Nicholas S. Ialongo
Center for Prevention and Early Intervention and Bloomberg School of Public Health, Johns Hopkins University

This study examined the longitudinal effects of 2 first-grade universal preventive interventions on academic outcomes (e.g., achievement, special education service use, graduation, postsecondary education) through age 19 in a sample of 678 urban, primarily African American children. The classroom-centered intervention combined the Good Behavior Game (H. H. Barrish, Saunders, & Wolfe, 1969) with an enhanced academic curriculum, whereas a second intervention, the Family–School Partnership, focused on promoting parental involvement in educational activities and bolstering parents' behavior management strategies. Both programs aimed to address the proximal targets of aggressive behavior and poor academic achievement. Although the effects varied by gender, the classroom-centered intervention was associated with higher scores on standardized achievement tests, greater odds of high school graduation and college attendance, and reduced odds of special education service use. The intervention effects of the Family–School Partnership were in the expected direction; however, only 1 effect reached statistical significance. The findings of this randomized controlled trial illustrate the long-term educational impact of preventive interventions in early elementary school.

Keywords: academic achievement, prevention and early intervention, educational outcomes, high school graduation, randomized controlled trial

As a result of federal policies such as the No Child Left Behind Act and the Individuals With Disabilities Education Act, there is an increasing emphasis on the use of evidence-based programs in schools to prevent disruptive behavior problems and promote academic success. Implementation of evidence-based, universal preventive interventions that simultaneously teach prosocial behavior and academic skills can assist schools in promoting healthy academic and social development among all students (Greenberg et al., 2003; Gresham, 2004; Walker, Ramsay, & Gresham, 2004). Yet there are relatively few educational and social–emotional preventive interventions that have been shown through rigorously designed randomized controlled trials to produce long-term educational outcomes (Catalano, Berglund, Ryan, Lonczak, & Hawkins, 2002; Durlak & Wells, 1997).

The need for efficacious programs is particularly great in urban communities, where the risk for school failure and early school leaving is considerably increased (Institute of Education Sciences, 2007; Perie, Grigg, & Donahue, 2005). The current study examined the longitudinal impacts of two first-grade universal preventive interventions targeting the early antecedent risk behaviors of poor academic achievement and aggressive/disruptive behavior and their distal correlates in a sample of urban, primarily African American children. Although both prevention programs were aimed to promote a range of outcomes related to classroom learning, one program focused on the proximal influence of the classroom environment, whereas the other program focused on the family–school partnership. By increasing understanding of the academic outcomes associated these types of universal preventive interventions, the current study could inform both policy and practice related to the future use of family- versus classroom-focused universal prevention programs in early elementary school.

Background and Theoretical Framework

Consistent with a three-tiered public health approach to prevention, *universal* preventive interventions target the general public or a whole population that has not been identified on the basis of individual risk (Mrazek & Haggerty, 1994). These programs are positive and proactive and are provided regardless of the student's risk status. A social–emotional learning curriculum delivered to an entire classroom is an example of a universal prevention program.

Catherine P. Bradshaw and Jessika H. Zmuda, Center for the Prevention of Youth Violence and Bloomberg School of Public Health, Johns Hopkins University; Sheppard G. Kellam, Bloomberg School of Public Health, Johns Hopkins University; Nicholas S. Ialongo, Center for Prevention and Early Intervention and Bloomberg School of Public Health, Johns Hopkins University.

This research was supported by National Institute of Mental Health Grant MH57005-02A and National Institute on Drug Abuse Grants NIDA RO1 DA11796-01A1 and P30MH06624 to Nicholas S. Ialongo. The writing of this article was supported by Centers for Disease Control and Prevention Grant K01CE001333-01 to Catherine P. Bradshaw.

Correspondence regarding this article should be addressed to Catherine P. Bradshaw, Department of Mental Health, Bloomberg School of Public Health, Johns Hopkins University, 624 North Broadway, Baltimore, MD 21205. E-mail: cbradsha@jhsph.edu

In contrast, *selective* preventive interventions target individuals or subgroups that are at elevated risk of developing disorders as a result of biological or social risk factors. Finally, *indicated* preventive interventions target individuals who are identified as having prodromal symptoms of severe behavioral problems or other disorders and whose symptoms are not yet serious enough to meet diagnostic criteria for a mental health or behavioral disorder (Mrazek & Haggerty, 1994). A similar tiered preventive intervention framework is utilized in positive behavior support (Sugai & Horner, 2006) and response to intervention (Hawken, Vincent, & Schumann, 2008).

A number of empirical studies support the focus on academic achievement and aggressive/disruptive behavior problems of preventive interventions for improving mental health, behavioral, and educational outcomes among urban children (Ialongo et al., 2006; Kellam, Mayer, Rebok, & Hawkins, 1998; McIntosh, Horner, Chard, Boland, & Good, 2006). For example, several studies have shown that learning problems predict mental health problems and anxiety and depressed mood in particular (Kistner, David, & White, 2003; Schwartz, Gorman, Duong, & Nakamoto, 2008). Similarly, aggressive behavior, displayed as early as first grade, has been shown to predict later substance use, antisocial behavior, and criminality (Schaeffer et al., 2006; Schaeffer, Petras, Ialongo, Poduska, & Kellam, 2003). Further complicating this association is the finding that aggressive/disruptive behavior problems often co-occur with poor academic achievement (Bradshaw, Buckley, & Ialongo, 2008; Herman & Ostrander, 2007), with some indication that conduct problems precede academic problems (Smart, Sanson, & Prior, 1996).

The life course/social field framework, originally described by Kellam, Branch, Agrawal, and Ensminger (1975) and more recently by Kellam and Rebok (1992), provided the conceptual and theoretical model for the design of the two universal preventive interventions examined in the current study. At the core of this framework is the concept that psychological well-being is reciprocally and positively related to how successfully people meet the social task demands faced at each stage of life. Success at an earlier stage of development may increase the likelihood of success at a later stage in the life course (Ialongo et al., 2006). Social task demands associated with the transition to first grade, the developmental stage on which the two preventive interventions focused, include academic achievement, compliance, attention, and participation in classroom and peer activities. The underlying theory of change (Izzo, Connell, Gambone, & Bradshaw, 2004) is that success in meeting these earlier social task demands will be associated with increased psychological well-being, academic outcomes, and overall success in meeting future task demands.

According to the theory, programs that target disruptive behavior problems and academic factors in first grade will improve academic outcomes on both the individual (student) and the collective (classroom) level (Ialongo et al., 2006; Kellam & Rebok, 1992). As a result of this collective classroom-level achievement, there should be a greater number of academically successful youths in the classroom for their classmates to model, and this would increase opportunities for learning, academic effort, and individual achievement (Bandura, Barbaranelli, Caprara, & Pastorelli, 1996). Success in the academic domain over the elementary school years would in turn reduce the child's likelihood of experiencing developmentally incongruent events, such as school dropout (Jessor, Turbin, & Costa, 1998; Patterson, DeBaryshe, & Ramsey, 1989; Patterson, Reid, & Dishion, 1992).

The extant theoretical and empirical work suggests that universal preventive interventions implemented in first grade would have a significant and long-term impact on students' academic performance and would reduce the risk for other untoward outcomes (Kellam et al., 1998). Therefore, a universal preventive intervention was designed for teachers and a second program was designed for parents; both interventions were to be implemented in first grade. Consistent with the life course/social field framework (Kellam & Rebok, 1992), the focus of both programs was on several outcomes specific to classroom learning and the management of disruptive behavior problems. Whereas the classroom-focused intervention was meant to enhance teachers' instructional and behavior management practices, a separate universal family-focused program was aimed at improving parental support and involvement in the academic process, as well as management of disruptive behavior problems. We theorized that both programs would reduce the early antecedent risk behaviors of poor academic achievement and disruptive behavior problems, on the individual and collective levels, by improving teachers' and parents' disciplinary practices and focus on academics (Ialongo et al., 2006; Kellam & Rebok, 1992). The current study examined the impact of the two programs on the students' educational outcomes in high school.

The Classroom-Centered and Family–School Partnership Universal Preventive Interventions

The first program, which we refer to as the classroom-centered (CC) intervention (Ialongo, Werthamer, Brown, Kellam, & Wai, 1999), is an integrated protocol designed to address both sets of early risk behaviors (poor achievement and aggressive and/or shy behavior). The CC intervention represents a combination of a behaviorally focused classroom management program called the Good Behavior Game (Barrish et al., 1969), which in previous trials demonstrated a beneficial impact on aggressive and shy behavior but not on achievement, with a set of specific enhancements to the academic curriculum, which in previous trials demonstrated a beneficial impact on achievement but only a modest crossover effect on aggression (Kellam et al., 1998). The CC intervention was designed to reduce the early risk behaviors of poor achievement and aggressive and shy behaviors through enhancements to the curriculum, improvements in teacher instructional and classroom behavior management practices, and specific strategies for children who were not performing adequately.

In addition to testing the CC intervention, we developed a second intervention, the Family–School Partnership (FSP; Canter & Canter, 1991; Ialongo et al., 1999), to improve collaboration between parents and teachers or school mental health professionals by enhancing parents' teaching and behavior management skills. The decision to develop and test a family-based universal preventive intervention was consistent with the life course/social field framework, which highlights the significance of parenting behaviors (e.g., monitoring, communication) as regulating influences on children's development. The FSP intervention was intended to help parents facilitate their children's early successful social adaptation in response to developmental challenges, which should in turn promote later successful adaptation when the children are

faced with new social task demands (Kellam & Rebok, 1992). Similarly, considerable empirical evidence has emerged documenting the important influences exerted by families on their children's academic success (Gallagher, 1987; Scott-Jones, 1984) and social development (Patterson et al., 1992), along with the benefits to children's behavior and achievement of strong parent–teacher partnerships and parental involvement (Epstein, 1983). The test of a family-based universal preventive intervention is consistent with the work of Hawkins et al. (1992) and Reid, Eddy, and Fetrow (1999), which demonstrated both the feasibility and the effectiveness of a universal family intervention implemented through the school.

Prior Research on the Impact of the CC and FSP Interventions

The two interventions were implemented during the 1993–1994 school year in first-grade classrooms in Baltimore City and were evaluated with a randomized controlled design (see Method section for additional information regarding the study design). Data were collected preintervention, and multi-informant data on behavioral and educational outcomes were collected in elementary, middle, and high school. At the end of the first and second grades, the CC intervention was shown to have significantly improved both academic achievement (reading and math) and classroom behavior (e.g., concentration problems, aggression, shyness). At the end of the sixth grade, children assigned to the CC intervention were significantly less likely to meet the criteria for a diagnosis of conduct disorder, to have been suspended from school, and to have received or been judged in need of mental health services in middle school by a teacher and/or parents (Ialongo, Poduska, Werthamer, & Kellam, 2001). At the end of the seventh grade, the CC intervention was shown to have significantly reduced the risk of initiating use of tobacco (Storr, Ialongo, Anthony, & Kellam, 2002). During Grades 6 through 8, the CC intervention was shown to have significantly reduced the risk of initiating use of hard drugs, such as heroin, crack, and cocaine powder (Furr-Holden, Ialongo, Anthony, Petras, & Kellam, 2004). Relative to the CC intervention, the FSP intervention has demonstrated a narrower range of significant impacts. In particular, none of the substance abuse or behavioral health effects of the FSP intervention proved statistically significant during middle school, with the exception of the effects for the FSP girls, who were significantly less likely to have been suspended than were control girls in sixth grade (Ialongo et al., 2001).

Overview and Rationale for the Current Study

The vast majority of the research on the CC and FSP has focused on behavioral outcomes through middle school. Less research has focused on longitudinal educational impacts of the two interventions. We aimed to add to the growing body of research documenting the positive outcomes of the CC and FSP universal preventive interventions by examining the distal impacts on academic outcomes by the end of high school. In the current study, we examined how exposure to either the CC or the FSP intervention in first grade affected standardized test performance in high school, teacher-rated academic achievement, special education service use, high school graduation, and college attendance. We theorized that the proximal impacts of the programs on reduced behavioral problems (Ialongo et al., 1999, 2001) would provide greater opportunity for learning (Scott & Barrett, 2004), which in turn would lead to improved academic performance, a reduced need for educational services (e.g., special education), and higher rates of high school graduation.

This study is significant in part because, to our knowledge, there are no randomized controlled trials of early elementary school-based, universal preventive interventions targeting early learning and behavior with follow-up from first grade through the end of high school. We are in the unique position of determining whether these early and relatively short-lived school-based interventions yielded long-term effects on a range of educational outcomes up to 12 years later. Moreover, because the study is based on an epidemiologically defined sample of urban, primarily economically disadvantaged African American children, such research could inform the use of preventive interventions in early elementary school to promote positive educational outcomes among urban African American youths.

Method

Study Design

The trial was designed to contrast the classroom-focused CC intervention with the family-based FSP intervention and a (third) control condition. This would allow us to determine which prevention program was more effective at promoting positive academic and mental health outcomes for the participating students. A randomized block design was employed, with schools serving as the blocking factor. Three first-grade classrooms in each of the nine urban elementary schools participated in the trial. Each classroom (including the teachers and students) within a school was randomly assigned either to one of the two intervention conditions (i.e., FSP or CC) or to the control condition, so that all three conditions were represented within each of the nine schools. We randomly assigned classrooms to one of the three conditions and balanced for student gender. Following an initial baseline assessment during the fall of the first-grade year, the interventions were implemented over the course of that first-grade school year. The CC and FSP interventions were implemented only in first grade, and there was no overlap in program content between the CC and FSP programs. The control classrooms followed the standard curriculum. Student participants were followed from first grade through age 19, and periodic assessments were conducted in Grades 1–3 and Grades 6–12 and at age 19 regarding a variety of mental health and academic outcomes.

Participants

Initial recruitment of first graders. In the fall of 1993, students from 27 classrooms in nine Baltimore City public elementary schools (chosen as representative of public elementary classrooms and schools in Baltimore City) were recruited to participate in a longitudinal study of the CC and FSP prevention programs. Recruitment began when the target sample was in kindergarten. Project staff attended parent–teacher meetings and distributed information regarding the project. When the target cohort entered the first grade, the project staff led an information session for parents

and attended parent–teacher conferences. Two parent liaisons were hired by the project to conduct follow-up visits to the parents in order to provide additional information about the project and obtain written parental consent. No incentives for initial participation were provided. Written informed consent was obtained from parents, and verbal assent was obtained from the youths. The recruitment procedure and all aspects of the data collection were approved by the institutional review board of Johns Hopkins University.

Written parental consent was obtained for 97% of the 678 children available for assessment in the fall of first grade. Three percent of the parents or guardians either refused to allow their children to participate in the assessments or failed to respond to the consent request. Chi-square analyses and *t* tests failed to reveal any significant differences in terms of sociodemographic characteristics (ethnicity, age, gender, and free lunch status) between the children for whom parental consent was obtained and those for whom it was not. The resulting epidemiologically-defined sample was 53% male and 86.8% African American. At the beginning of first grade, the sample of students ranged in age from 5.3 to 7.7 years (M = 6.2 years, SD = 0.34). The majority of children (68.3%) received free or reduced-priced lunch. Additional demographic characteristics of the sample by intervention status are reported in Table 1.

Follow-up data collection. Data on a number of education performance indicators (e.g., standardized test performance, high school graduation) were collected for the current study in high school (Grades 6–12) and at age 19. Written consent was required from those participants who were 18 or older. Of those turning 18 during the spring fielding period, 192 (28.3%) gave written consent to participate. Written parental consent for participation was obtained for 382 (56.3%) of the 678 youths who had yet to turn 18. A total of 574 youths (84.7%) consented to a Grade 12 assessment. Parents refused consent for 39 youths (5.7%) and 9 of those participants who were 18 or older refused, for a total of 48 refusals (7.1%). Three participants were deceased (0.4%); the remaining 53 youths (7.8%) are currently being located to obtain consent for future data collections. No significant differences have been found in attrition or refusal rates between or across intervention conditions.

Table 1
Baseline Characteristics of the Participants by Intervention Condition and Gender

Baseline characteristic	Overall	Boys	Girls
Age, *M* (*SD*)			
CC	6.20 (0.34)	6.21 (0.33)	6.18 (0.35)
FSP	6.25 (0.37)	6.29 (0.43)	6.20 (0.29)
Control	6.25 (0.36)	6.26 (0.38)	6.24 (0.33)
% receiving free lunch			
CC	68.4	75.2	60.2
FSP	67.4	62.3	67.3
Control	71.0	73.6	75.2
% African American			
CC	87.2	88.3	85.8
FSP	83.8	82.6	85.3
Control	83.5	80.3	86.8
Early academic readiness, *M* (*SD*)			
CC	3.14 (1.35)	3.34 (1.30)	2.89 (1.38)
FSP	2.85 (1.34)	3.03 (1.34)	2.65 (1.33)
Control	2.57 (1.36)	2.80 (1.43)	2.32 (1.26)

Note. CC = classroom-centered intervention; FSP = Family–School Partnership intervention; early academic readiness = teacher's rating in fall of first grade, with higher scores indicating less readiness.

Interventions

The CC intervention. The CC intervention was designed to reduce the early risk behaviors of poor achievement and aggressive and shy behaviors by enhancing the classroom curriculum and teacher instructional and behavior management practices. In particular, the CC intervention featured (a) curricular enhancements, (b) improved classroom behavior management techniques, and (c) accompanying strategies for children not responding adequately to the universal intervention. To increase listening and comprehension skills, the reseachers augmented the intervention with an interactive, read-aloud component. Journal writing activities and the Reader's Theater, a dramatic presentation of written work in a script form that includes expressive voices and gestures, were added to improve compositional and reading skills. These activities were intended to make the core curriculum more meaningful and fun for the students. To improve critical thinking skills, the researchers incorporated a new component called Critique of the Week. This directed-thinking activity was developed to help students learn strategies for analyzing perspectives. It uses the context of images and resources from the students' daily life to teach students to examine the content, to look at the way they think, and to formulate their own position with a system of value and reasoning. The Mimosa math curriculum augmented the existing math curriculum and featured a whole-language approach to promoting math skill development. The academic component of the CC intervention divided the class into small, diverse groups, which provided the underlying structure for the curricular and behavioral components of the intervention.

The researchers augmented the existing classroom behavior management practices with the Good Behavior Game (Barrish et al., 1969). As briefly described in the Introduction, the Good Behavior Game is a whole-class strategy intended to decrease disruptive behaviors. Children are assigned to teams, and only those teams that do not exceed a specified criterion of precisely defined off-task, disruptive, and aggressive behaviors are allowed to "win." Teachers were given additional strategies to use with children who failed to respond to the Good Behavior Game and/or the curricular enhancements. The strategies employed with respect to academic nonresponders included individual or small-group tutoring and modifications in the curriculum to address individual learning styles.

The FSP intervention. As briefly described in the Introduction, the FSP (Canter & Canter, 1991; Ialongo et al., 1999) was developed to improve collaboration between parents and teachers and school mental health professionals and to enhance parents' teaching and behavior management skills. The major features of the FSP intervention are (a) training teachers, school mental health professionals, and other relevant school staff in parent–school communication and partnership building (Canter & Canter, 1991); (b) weekly home–school learning and communication activities; and (c) a series of nine workshops for parents led by a first-grade teacher and a school psychologist or social worker. The workshop

series for parents began immediately after the pretest assessments in the fall of the first-grade school year and ran for 7 consecutive weeks (one per week) through early December. Two follow-up or booster workshops were held during the winter and spring semesters.

The initial parent workshops were intended to establish an effective and enduring partnership between parents and school staff, and they set the stage for parent–school collaboration to support children's learning and behavior. In the first workshop, called Read Aloud, teachers shared with parents the benefits of reading aloud to their children along with strategies to enhance the experience. Parents were loaned a different book each week to read aloud to their child; the books contained sample questions and activities developed by Karweit and Bond (1993) and Handel and Goldsmith (1990) that have been found to promote development of literacy skills (Whitehurst, Epstein, Payne, Crone, & Fischel, 1994). The second workshop focused on Fun Math activities that were derived from the University of California at Berkeley's Family Math program. These activities have been shown to be effective in stimulating children's understanding of mathematical concepts and operations (Stenmark, Thompson, & Cossey, 1986). During this workshop, parents were given a kit of "manipulatives" to use when carrying out the weekly Fun Math activities sent home by their child's first-grade teacher.

The next five workshops focused on effective disciplinary strategies. The Parents and Children series, a videotape modeling and group discussion program (Webster-Stratton, 1984), formed the basis for the positive discipline component of the FSP intervention. The series was led by the school psychologist or social worker and covered topics that included effective praise, play, limit setting, time-out versus spanking, and problem solving. Parents observed a series of videotapes of modeled parenting skills. After viewing each vignette, the leader paused the videotape and asked open-ended questions about the scenes. Parents reacted to and discussed the episodes and problem solved alternative approaches. Many situations were role-played and rehearsed by group members. Families were also asked to discuss and problem solve other problem situations that occur at home. The facilitating school psychologist/social worker also established a voicemail system with which to maintain parent involvement and provide consultation as needed with respect to learning or behavior management difficulties. To further foster family–school communication, researchers asked parents to fill out and return comment sheets indicating whether they had completed the assigned weekly home learning activities and if they had encountered any problems in doing so.

Intervention Fidelity

The training and intervention manuals were precisely and uniformly delineated and codified, and the content of training and intervention contacts was standardized. This allowed us to monitor and sustain the integrity of the first-grade interventions. In addition, each intervener had available a number of materials designed to foster the correct execution of the interventions (e.g., detailed outlines and checklists that prescribed the necessary materials for each intervention contact, the specific themes or tasks that needed to be covered). Finally, the intervener had extensive training prior to the initiation of the interventions and received ongoing supervision, feedback, and training throughout the entire intervention period. Teachers who took part in the CC received 60 hr of training and direct supervision in its use. Parents who participated in the FSP were offered a total of nine workshop sessions, each of which was approximately 90 min in length. The monitoring of fidelity of implementation for the CC intervention involved three parts: (a) measures of setting up the classroom, (b) independent classroom observations, and (c) classroom visit record reviews. For the FSP intervention, interveners were required to provide documentation of each contact with parents (e.g., workshop attendance, level of parental participation, parental and student compliance with homework assignments).

Measures

Student demographic information. The school district provided information on the students' gender, ethnicity, and free or reduced meals status in first grade.

Externalizing behavior problems. The Teacher Observation of Classroom Adaptation—Revised (TOCA–R; Werthamer-Larsson, Kellam, & Wheeler, 1991) was administered in the fall of the first grade to the students' teachers. The TOCA–R is a brief measure of the adequacy of each child's performance on core tasks in the classroom. Using a structured interview conducted by a trained member of the project staff, the teachers rated each child's behavior on a scale measuring frequency of occurrence from 1 (*almost never*) to 6 (*almost always*). Academic readiness ratings in the fall of first grade was a covariate in the analyses in the current study. This subscale comprises 9 items (e.g., "concentrates," "pays attention," "is eager to learn," "learns up to ability," "works hard," "stays on task"). The items were averaged to create a score from 1 to 6; higher scores indicated less academic readiness. Prior research on the TOCA has indicated that the Academic Readiness subscale is internally consistent (Cronbach's $\alpha = .97$; Werthamer-Larsson et al., 1991), and the scores are predictive of subsequent externalizing behavior problems, such as adjudication for a violent crime in adolescence and meeting the criteria for a diagnosis of antisocial personality disorder at age 19 (Petras, Chilcoat, Leaf, Ialongo, & Kellam, 2004; Schaeffer et al., 2003).

Kaufman Test of Educational Achievement (Grade 12, Reading and Math). The Kaufman Test of Educational Achievement (KTEA; Kaufman & Kaufman, 1985) was developed to assess students' school achievement in Grades 1–12. The brief test administered in the current study assessed academic ability in the areas of reading and math. The measure is widely used, and the scores have strong internal consistency (split-half reliability coefficients for each age-group ranged from .85 to .95) and test–retest reliability (coefficients ranged from .83 to .97; Worthington, 1987). Prior research has shown the KTEA scores to be correlated with other commonly used achievement tests (e.g., Wide Range Achievement Test, Peabody Individual Achievement Test, Metropolitan Achievement Test, Stanford Achievement Test, Comprehensive Test of Basic Skills, and the Kaufman Assessment Battery for Children; Worthington, 1987).

Teacher-rated academic performance (Grades 6–12). The Teacher Report of Classroom Behavior Checklist (Ialongo et al., 2001) was used in Grades 6–12 to assess both classroom behavior and academic performance. This measure is an adaptation of the TOCA–R (described above). Teachers in Grades 6–12 responded

to a question on the Teacher Report of Classroom Behavior Checklist regarding the youth's academic performance in class on the following 5-point scale: *excellent*, *good*, *fair*, *barely passing*, or *failing*. An average of the ratings over the 6 years was computed and used as a composite academic performance indicator; higher scores indicated better performance.

Special education service use (Grades 1–12). The school district provided official records of special education use. These records included students who had an individualized education program. In some cases (less than 5% of the sample), data from the district records were missing on this variable. In such cases, the teacher-reported special education service use was analyzed.

High school graduation. Data were obtained from the district to determine whether the participant had graduated from high school or had passed the General Educational Development test (approximately 4% of the sample). In some cases (less than 5% of the sample), data from the school records were missing on this variable. In such cases, the self-report data from the student's age 19 interview were used.

College attendance. At the age 19 interview, the youths were asked whether they had attended college (e.g., 4-year college, junior college).

Overview of Statistical Analyses

We used mixed-model regression analysis (i.e., random effects regression; Gibbons, Hedeker, Watemaux, & Davis, 1988) to evaluate the impact of the two classroom-level interventions on the educational outcomes. The first-grade classroom was included as a random effect in the analyses (which accounted for the clustering of students within the 27 participating first-grade classrooms), and intervention condition was modeled as a fixed effect (Gibbons et al., 1988). Dummy variables were created to allow for planned contrasts between the CC intervention and control condition and the FSP intervention and control condition. Consistent with prior research, the analyses were conducted with the full sample and were stratified by gender. All of the analyses examined the impact of the interventions after adjusting for the baseline (the fall of first grade) level of academic readiness as rated by teachers. Cohen's *d* (1992) effect size estimates were computed for continuous outcomes, and odds ratios were computed for dichotomous outcomes.

Preliminary Analyses

Equivalence of the intervention conditions at baseline. Chi-square analyses and analyses of variance revealed that the intervention conditions were equivalent with respect to child age, gender, ethnicity, free lunch status, achievement levels, and parenting practices at baseline (i.e., fall of first grade; Ialongo et al., 1999). A significant difference ($p < .05$) was found between the CC intervention and controls in terms of teacher ratings of early academic readiness; therefore, this variable was included as a covariate in the regression analyses.

Attrition analyses. Of the 653 children with consent to participate in the evaluation in the fall of first grade, 597 (i.e., 91.3%) completed the fall and spring of first-grade assessments and remained in their assigned intervention condition over the first-grade year. A total of 574 students (84.7%) completed assessments during the spring of 12th grade. At age 19, 541 (79.8%) consented to participate. The response rate has been consistently high through the age 19 follow-up, as illustrated by the 80% or greater follow-up rate in the annual assessments. A 20% loss to follow-up is considered a borderline situation, given that inferences could be affected by this amount of missing data. This would be particularly true if there were systematic loss to follow-up (Brown, 1992); however, we have not found evidence of such systematic loss through the age 19 assessments. As noted above, there were no significant differences between the intervention conditions in terms of rates of attrition at the 12th-grade follow-up. Furthermore, there were no differences in the sociodemographic characteristics (ethnicity, gender, age, or free lunch status) in terms of rates of attrition at 12th grade or at the age 19 interview across the intervention conditions.

Level of participation/implementation in the CC and FSP interventions. Each of the nine CC intervention classrooms was assigned a score from 0 to 100 that represented the percentage of the teacher's implementation of the intervention as designed. Scores were based on the three sources of implementation data identified previously: (a) measures of setting up the classroom, (b) independent classroom observation sessions, and (c) reviews of classroom visit records. CC implementation scores ranged from 30% to 78% ($Mdn = 64.37\%$, $M = 59.9\%$, $SD = 17.03\%$). All but two of the nine CC intervention teachers implemented more than 50% of the intervention protocol. Parents or caregivers in the FSP intervention attended an average of 4.02 (range 0–7, $Mdn = 5.0$, $SD = 2.38$) of the 7 core parenting sessions offered in the fall of first grade, or 57.4% of the available sessions. Just under 13% (12.7 %) of the parents or caregivers failed to attend any of the core workshops, whereas just over one third (35.3%) of the parents attended at least 6 of the 7 sessions. On average, parents completed 39.15 ($SD = 16.54$) of the 64 activities (i.e., 60.9%) of weekly take-home Read Aloud and Fun Math activities. Approximately one third (35.7%) of the families in the FSP condition completed 75% or more of the activities, whereas only 2.3% of the families failed to complete all of the activities.

Results

Effects of the CC Intervention

The mixed-model regression analyses indicated a significant effect of the CC intervention on KTEA reading performance in the overall sample (i.e., for boys and girls, $B = 1.828$, $p < .01$; for boys, $B = 2.43$, $p < .05$; see means for the full sample and by gender in Table 2 and regression results in Table 3). A similar significant CC intervention effect was found for KTEA math performance both in the overall sample ($B = 3.57$, $p < .01$) and for the boys ($B = 4.27$, $p < .01$), whereas a marginally significant effect was observed for girls ($B = 2.81$, $p = .072$). We observed a marginally significant CC intervention effect in the overall sample ($B = 0.168$, $p = .081$) for teacher-rated academic performance in Grades 6–12 (averaged), such that the children in the CC condition tended to have their academic performance rated more favorably than did control children. We also observed a significant effect of the CC intervention on special education service use in the overall sample ($B = -0.705$, odds ratio [OR] = .494, $p < .05$) and for the boys in the CC condition ($B = -0.934$, OR = .393, $p < .01$; see percentages in Table 2 and regression results in Table 4). That

Table 2
Adjusted Descriptive Analyses on Educational Outcomes Among the Participants by Intervention Condition and Gender

	Overall		Boys		Girls	
Educational outcome	*M*	*SE*	*M*	*SE*	*M*	*SE*
Grade 12 KTEA–Reading						
CC	42.624	0.392	43.100	0.566	42.178	0.536
FSP	42.299	0.422	42.287	0.669	42.313	0.639
Control	40.900	0.416	40.879	0.608	40.895	0.554
Grade 12 KTEA–Math						
CC	43.671	0.616	43.705	0.813	43.520	0.945
FSP	42.739	0.639	42.653	0.911	42.789	1.131
Control	40.102	0.655	39.331	0.877	40.920	0.977
Academic performance						
CC	3.013	0.054	2.770	0.072	3.269	0.077
FSP	2.993	0.052	2.827	0.071	3.174	0.084
Control	2.842	0.055	2.639	0.074	3.095	0.077
Special education service, %						
CC	22.4	2.8	24.6	3.9	19.5	4.0
FSP	29.3	3.1	34.2	4.4	24.0	4.0
Control	34.8	2.9	43.2	4.3	26.2	3.8
High school graduation, %						
CC	61.9	3.3	53.3	4.5	73.6	4.9
FSP	57.3	4.6	56.1	4.6	58.4	6.2
Control	51.3	3.4	44.8	5.0	58.6	4.6
College attendance, %						
CC	31.5	2.9	26.0	3.6	40.3	4.9
FSP	28.3	3.8	22.4	4.1	34.5	6.3
Control	19.4	3.0	12.8	3.9	26.4	4.7

Note. All descriptive analyses adjusted for teacher-rated fall of first-grade academic readiness. Overall indicates the pooled sample for both boys and girls. KTEA–Reading = Kaufman Test of Educational Achievement Brief Reading Comprehension, administered in Grade 12; CC = classroom-centered intervention; FSP = Family–School Partnership intervention; KTEA–Math = Kaufman Test of Educational Achievement Math Calculation, administered in Grade 12; academic performance = teacher-rated academic performance, averaged across Grades 6–12; special education service = special education service use at any time in Grades 1–12; *SE* = standard error of the mean.

is, the odds of the children in the CC receiving special education at any point between Grades 1 and 12 were reduced nearly 50% for the overall sample as compared to the children in the control condition, whereas the odds of special education service use were reduced nearly 60% for the boys in the CC condition. No significant effects of the CC intervention were observed on special education service use among the girls. With regard to high school graduation, there was a significant positive effect of the CC intervention in the overall sample ($B = 0.532$, OR = 1.702, $p < .05$) or for the girls ($B = 0.750$, OR = 2.118, $p < .05$). Finally, with regard to college attendance, there was a significant effect of the CC intervention in the overall sample ($B = 0.798$, OR = 2.222, $p < .05$) and among the boys in the CC condition ($B = 0.982$, OR = 2.670, $p < .05$). A marginally significant intervention effect was observed among the girls in the CC condition ($B = 0.690$, OR = 1.993, $p = .052$).

Effects of the FSP Intervention

The mixed-model regression analyses indicated that there was a marginally significant effect of the FSP intervention on KTEA reading performance in the overall sample ($B = 1.160$, $p = .081$) but no significant effects for boys or for girls in the stratified analyses (see means for the full sample and by gender in Table 2 and regression results in Table 3). There was a significant effect on KTEA math performance in the overall sample ($B = 2.001$, $p < .05$) and a marginally significant effect for the boys ($B = 2.530$, $p = .059$). There were no significant effects of the FSP intervention on teacher-rated academic performance in Grades 6–12. Furthermore, there were no statistically significant effects of the FSP intervention on special education service use, high school graduation, or college attendance (see percentages in Table 2 and regression results in Table 4).

Discussion

In the current study, we used a randomized controlled trial design to examine the effects of two preventive interventions implemented in first grade on academic outcomes through age 19 in a sample of urban children. When we controlled for academic readiness in the fall of first grade, the longitudinal analyses indicated that the CC intervention, which comprised the Good Behavior Game and an enhanced academic curriculum, was associated with significant improvements in reading and general academic achievement, high school graduation, and college attendance, as well as with reductions in special education service use. The effect sizes of the CC intervention (as indicated by the Cohen's *d* and odds ratios) are well within the practical significance range according to Lipsey (1998) and are in the small-to-moderate range according to Cohen (1992). The effect sizes varied somewhat by gender, such that the effects tended to be greatest in the overall sample and for boys, whereas the only significant effect among girls was on high school graduation.

Table 3
Summary of Regression Analyses Examining Intervention Impact on Continuous Educational Outcomes Through Age 19

Continuous outcome	*B*	*SE*	*p*	Cohen's *d*
		Grade 12 KTEA–Reading		
Overall				
CC	1.828**	0.617	.009	0.319
FSP	1.160†	0.624	.081	0.252
Boys				
CC	2.433*	0.915	.017	0.394
FSP	1.296	0.981	.204	0.232
Girls				
CC	1.281	0.927	.185	0.253
FSP	1.039	0.924	.276	0.254
		Grade 12 KTEA–Math		
Overall				
CC	3.569**	0.915	.001	0.420
FSP	2.001*	0.848	.030	0.308
Boys				
CC	4.273**	1.328	.005	0.540
FSP	2.530†	1.251	.059	0.391
Girls				
CC	2.809†	1.465	.072	0.290
FSP	1.508	1.581	.354	0.192
		Academic performance		
Overall				
CC	0.168†	0.090	.081	0.225
FSP	0.116	0.071	.119	0.200
Boys				
CC	0.118	0.110	.295	0.174
FSP	0.148	0.096	.141	0.251
Girls				
CC	0.200	0.119	.113	0.240
FSP	0.067	0.122	.589	0.101

Note. The control group is dummy coded as the base group in all analyses. All mixed-model regression analyses included teacher-rated fall of first grade academic readiness as a covariate. Overall indicates the pooled sample for both boys and girls. CC = classroom-centered intervention; FSP = Family–School Partnership intervention; *B* = unstandardized regression coefficient; *SE* = standard error of the estimate; KTEA–Reading = Kaufman Test of Educational Achievement Brief Reading Comprehension, administered in Grade 12; KTEA–Math = Kaufman Test of Educational Achievement Math Calculation, administered in Grade 12; academic performance = teacher-rated academic performance, averaged across Grades 6–12.
† $p < .10$. * $p < .05$. ** $p < .01$.

Although the overall effects of the CC intervention were significant for several outcomes, the stratified analyses indicated that many of the intervention effect sizes were larger for boys than for girls. It is possible that the targets of the CC intervention (e.g., impulsive and disruptive behavior, academic focus) were more relevant to long-term academic performance for boys than for girls. Alternatively, as noted in a previous study on the outcomes in elementary school, girls tended to display more shy behavior than did boys, whereas boys were at greater risk for displaying externalizing behavior problems (Ialongo et al., 1999). Therefore, it seems plausible that poorly achieving girls receive less attention from teachers than do poorly achieving boys. Researchers should examine these gender differences in greater detail to discern whether elements of the program could be modified or enhanced to address proximal factors associated with girls' long-term academic performance.

Despite research highlighting the significance of parenting factors on multiple aspects of children's development (Collins, Maccoby, Steinberg, Hetherington, & Bornstein, 2000; Patterson et al., 1989), there were relatively few significant academic outcomes associated with the FSP intervention. In particular, the FSP intervention was associated with statistically significant improvements in the overall sample's math performance and with marginally significant improvements in boys' math performance and the overall sample's reading performance. All other nonsignificant effects were, however, in the expected direction. Similar trends for the FSP intervention have been noted for other outcomes, such as criteria for conduct disorder, need for mental health services, teacher-rated problem behaviors, social participation/shy behavior (Ialongo et al., 1999, 2001), and illegal drug use (Furr-Holden et al., 2004). This finding suggests that further enhancement may be needed to optimize the overall impact of the FSP intervention (Patrikakou & Weissberg, 2007).

Another possible explanation for the stronger impact of the CC program than the FSP program on the education outcomes may be the differential amount of training and coaching the CC teachers

Table 4
Summary of Regression Analyses Examining Intervention Impact on Categorical Educational Outcomes Through Age 19

Categorical outcome	*B*	*SE*	Wald χ^2	*p*	Odds ratio
		Special education service			
Overall					
CC	−0.705	0.140	50.05	.013	0.494*
FSP	−0.180	0.197	46.78	.444	0.835
Boys					
CC	−0.934	0.141	38.10	.009	0.393**
FSP	−0.300	0.231	31.16	.337	0.741
Girls					
CC	−0.389	0.256	25.89	.302	0.678
FSP	−0.024	0.334	32.49	.944	0.976
		High school graduation			
Overall					
CC	0.532	0.439	27.67	.039	1.702*
FSP	0.202	0.345	21.25	.473	1.224
Boys					
CC	0.405	0.377	11.59	.108	1.499
FSP	0.435	0.427	8.19	.115	1.545
Girls					
CC	0.750	0.760	18.68	.037	2.118*
FSP	−0.044	0.359	16.47	.906	0.957
		College attendance			
Overall					
CC	0.798	0.750	39.83	.018	2.222*
FSP	0.434	0.467	17.94	.153	1.543
Boys					
CC	0.982	1.290	10.39	.042	2.670*
FSP	0.571	0.638	8.95	.113	1.770
Girls					
CC	0.690	0.709	18.71	.052	1.993†
FSP	0.397	0.677	14.77	.384	1.487

Note. The control group is dummy coded as the base group in all analyses. All mixed-model regression analyses included teacher-rated fall of first-grade academic readiness as a covariate. Overall indicates the pooled sample for both boys and girls. Special education service = special education service use at any point in Grades 1–12; CC = classroom-centered intervention; FSP = Family–School Partnership intervention; *B* = unstandardized regression coefficient; *SE* = standard error of the estimate.
† $p < .10$. * $p < .05$. ** $p < .01$.

and FSP parents received. Whereas the CC intervention teachers received 60 hr of training and direct supervision in the use of the programs, parents in the FSP program were offered a total of only nine 90-min workshop sessions. However, it was not feasible, from the standpoint of logistics or cost, to provide all first-grade parents with as much training as the CC teachers received. It is also possible that a universal family intervention may not yield an impact sufficient to justify the resources expended, and thus it may be more efficient to target family supports to the children at greatest risk (e.g., those not responding adequately to the universal CC intervention).

We did not test the combined effect of the CC and FSP interventions within the current trial. However, it is possible that the combination of the CC intervention and a parenting program that uses language similar to that of the CC intervention and focuses specifically on the skills fostered through the CC program could build on and help the children generalize the skills developed through the classroom-based program to other settings (e.g., home, neighborhood). This combination in turn might result in greater effects of the integration of CC and a family program than would either program in isolation. Or, as noted above, the family component could be employed for children who do not respond adequately to the CC universal model (see Sugai & Horner, 2006, for a description of multitiered preventive interventions).

The current study focused on the intent-to-treat impact of the CC and FSP interventions; however, future research should also take into consideration implementation level and parent participation as factors influencing program outcomes. For example, the level of parent participation in the FSP intervention varied across students. This is a common concern in parent-focused preventive interventions and services (McKay et al., 2004). In fact, prior research by Jo and Muthén (2002) indicated that the intervention effects of the FSP intervention on subsequent teacher ratings of the children's academic readiness were weaker among the children whose parents had the lowest participation in the FSP intervention. Further work should identify effective strategies for optimizing

parent involvement and engagement in similar preventive interventions (McKay et al., 2004).

Although the randomized controlled longitudinal design bolsters a potentially causal interpretation of the intervention effects, further exploration of the specific mechanisms mediating the change process is required in future studies (Kellam et al., 1998). As described previously, prior research on this sample indicates that the CC intervention is associated with reductions in the child's level of disruptive behavior in elementary and middle school (Ialongo et al., 1999, 2001). The behavior management component of the CC intervention (i.e., Good Behavior Game) may have provided children with a greater opportunity to learn (Scott & Barrett, 2004) by reducing behavior problems, whereas the academic curriculum component of the CC program may have enhanced the children's early academic skills (McIntosh et al., 2006). In contrast, the increased reliance on special education services among the children in the control condition may have influenced their long-term outcomes (e.g., increased dropout, school failure) as a result of labeling or other processes (Osterholm, Nash, & Kritsonis, 2007). An important area for future research is identification of the specific pathways or mechanisms through which the CC and FSP interventions influence the academic outcomes. Furthermore, the impact of the intervention may vary as a function of the quality of subsequent educational services or the academic curriculum, as well as the presence of other child, family, or community risk factors. Thus, potential moderators of intervention outcomes warrant further exploration.

Given the combined focus on behavior and academics in the CC intervention, it is unclear whether the Good Behavior Game or the academic components of the CC program account for the observed effects on educational outcomes. A previous randomized trial by Kellam and colleagues, which included a different sample of Baltimore City elementary school students, found that the Good Behavior Game intervention alone had significant effects on disruptive behavior problems but did not impact educational outcomes (Kellam et al., 2008; Kellam, Rebok, Ialongo, & Mayer, 1994). The findings of the current study, which tested a combination of the Good Behavior Game and the enhanced academic curriculum through the CC condition, suggest there may be a synergistic effect on education outcomes when the Good Behavior Game is combined with an enhanced educational program. The design of the current study, however, precludes us from determining the independent contributions of the Good Behavior Game and the enhanced academic curriculum on the observed outcomes.

Strengths of the current study include the focus on longitudinal outcomes (first grade through age 19) and the use of multiple indicators of academic performance. A relatively consistent pattern of findings emerged across the different outcomes, and this suggests that the intervention effects were not limited to select outcomes or one source of information (e.g., self-report). It is interesting that the only outcome for which a significant intervention effect did not emerge was the teacher-reported outcome of academic achievement. However, one effect approached statistical significance, and all the trends were in the expected direction. Additional research could help us understand the factors influencing teachers' ratings of children's academic performance.

Conclusions and Implications

The findings of the current study provide further evidence of the enduring effects of the yearlong CC intervention in first grade across a range of educational outcomes. The strong beneficial effects of the CC intervention on academic achievement, as measured by higher scores on standardized achievement tests, reduced utilization of special education services, higher rates of high school graduation, and higher rates of attending college, will likely translate into increased employment opportunities for these youths and concomitant reductions in the risk for mental and behavioral health problems (Garnier, Stein, & Jacobs, 1997; Prevatt & Kelly, 2003; Vernez, Krop, & Rydell, 1999). These findings are particularly noteworthy, given that they occurred within a relatively high risk population of urban, low-income, primarily African American children, for whom the risk of academic problems, special education service use, and school dropout is significantly increased (Institute of Education Sciences, 2007; Miller-Cribbs, Cronen, & Davis, 2002). These effects are also impressive, given that the CC intervention was of relatively low intensity and was administered over the course of a single academic year.

There is growing interest in economic analysis of school-based prevention programs in an effort to determine the fiscal impact of implementing the programs. Such work is particularly important given the current fiscal climate and increasing emphasis on high-quality implementation of evidence-based programs, which can require considerable resources in terms of time, money, and training (Gottfredson & Gottfredson, 2002). Applying a benefit–cost model to the effects of the Good Behavior Game on (only) the reduced use of tobacco observed in the first-generation trials (Kellam & Anthony, 1998), Aos, Lieb, Mayfield, Miller, and Pennucci (2004) concluded that the Good Behavior Game was associated with benefits of $25.92 for each dollar of program cost. This put the Good Behavior Game among the top five programs reviewed with respect to benefits per dollar of cost (Aos et al., 2004). The benefit–cost ratio is likely considerably greater when the broader range of behavioral, mental health, and educational outcomes is considered. The savings are likely even greater when the reduced reliance on special education, mental health, and juvenile justice services resulting from the 1-year intervention is considered. A cost–benefit analysis of the CC intervention on increased high school graduation is currently under way, and preliminary findings suggest that the combined classroom-based model is a worthwhile investment for schools and communities.

Taken together, the current findings highlight the significant impact of the relatively modest 1-year CC universal preventive intervention for promoting a range of positive educational outcomes among economically disadvantaged, urban African American youths. The findings suggest that early intervention with such youths can have a positive impact 12 years later. As noted above, this is the only randomized controlled study of which we are aware that has a follow-up of a sufficient length to attest to the positive consequences of targeting early academic achievement and behavior. Thus, the current findings bolster the available data indicating the effectiveness of the CC model on a broad range of long-term educational outcomes, and they represent an important and novel extension of our prior work on the behavioral outcomes of these universal preventive interventions.

References

Aos, S., Lieb, R., Mayfield, J., Miller, M., & Pennucci, A. (2004). *Benefits and costs of prevention and early intervention programs for youth: Technical appendix.* Olympia, WA: Washington State Institute for Public Policy (p. 97). Retrieved July 25, 2008, from http://wsipp.wa.gov/rptfiles/04-07-3901a.pdf

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Multifaceted impact of self-efficacy beliefs on academia functioning. *Child Development, 67,* 1206–1222.

Barrish, H. H., Saunders, M., & Wolfe, M. D. (1969). Good Behavior Game: Effects of individual contingencies for group consequences and disruptive behavior in a classroom. *Journal of Applied Behavior Analysis, 2,* 119–124.

Bradshaw, C. P., Buckley, J., & Ialongo, N. (2008). School-based service utilization among urban children with early-onset educational and mental health problems: The squeaky wheel phenomenon. *School Psychology Quarterly, 23,* 169–186.

Brown, C. H. (1992, November). Handling missing data in behavioral studies. In *Advanced methodology and statistics.* Seminar conducted at the 26th Annual Convention of the Association for Advancement of Behavioral Therapy, Boston, MA.

Canter, L., & Canter, M. (1991). *Parents on your side: A comprehensive parent involvement program for teachers.* Santa Monica, CA: Canter.

Catalano, R. F., Berglund, M. L., Ryan, J. A. M., Lonczak, H. S., & Hawkins, J. D. (2002). Positive youth development in the United States: Research findings on evaluations of positive youth development programs. *Prevention & Treatment, 5,* Article 15. Retrieved from http://journals.apa.org/prevention/volume5/pre0050015a.html

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Collins, W. A., Maccoby, E. E., Steinberg, L., Hetherington, E. M., & Bornstein, M. H. (2000). Contemporary research on parenting: The case for nature and nurture. *American Psychologist, 55,* 218–232.

Durlak, J. A., & Wells, A. M. (1997). Primary prevention mental health programs for children and adolescents: A meta-analytic review. *American Journal of Community Psychology, 25,* 115–152.

Epstein, J. L. (1983). *Effects on parents of teacher practices of parent involvement* (Report No. 346). Baltimore, MD: Center for Social Organization of Schools, Johns Hopkins University.

Furr-Holden, C. D. M., Ialongo, N., Anthony, J. C., Petras, H., & Kellam, S. G. (2004). Developmentally inspired drug prevention: Middle school outcomes in a school-based randomized prevention trial. *Drug and Alcohol Dependence, 73,* 149–158.

Gallagher, J. J. (1987). Public policy and the malleability of children. In J. J. Gallagher & C. T. Ramey (Eds.), *The malleability of children* (pp. 199–208). Baltimore, MD: Brookes.

Garnier, H. E., Stein, J. A., & Jacobs, J. K. (1997). The process of dropping out of high school: A 19-year perspective. *American Educational Research Journal, 34,* 395–419.

Gibbons, R., Hedeker, D., Watemaux, C., & Davis, J. (1988). Random regression models: A comprehensive approach to the analysis of longitudinal psychiatric data. *Psychopharmacology Bulletin, 24,* 438–443.

Gottfredson, D. C., & Gottfredson, G. D. (2002). Quality of school-based prevention programs: Results from a national survey. *Journal of Research on Crime and Delinquency, 39,* 3–35.

Greenberg, M. T., Weissberg, R. P., O'Brien, M. U., Zins, J. E., Fredericks, L., Resnik, H., & Elias, M. J. (2003). School-based prevention: Promoting positive social development through social and emotional learning. *American Psychologist, 58,* 466–474.

Gresham, F. M. (2004). Current status and future directions of school-based behavioral interventions. *School Psychology Review, 33,* 326–343.

Handel, R., & Goldsmith, E. (1990). *Family reading: A home–school partnership.* Montclair, NJ: Montclair State College.

Hawken, L. S., Vincent, C. G., & Schumann, J. (2008). Response to intervention for social behavior: Challenges and opportunities. *Journal of Emotional and Behavioral Disorders, 16,* 213–225.

Hawkins, J. D., Catalano, R., Morrison, D., O'Donnell, J., Abbott, R. D., & Day, L. E. (1992). The Seattle Social Development Project: Effects of the first four years on protective factors and problem behaviors. In J. McCord & R. E. Tremblay (Eds.), *Preventing antisocial behavior: Interventions from birth through adolescence* (pp. 139–161). New York: Guilford Press.

Herman, K. C., & Ostrander, R. (2007). The effects of attention problems on depression: Developmental, academic, and cognitive pathways. *School Psychology Quarterly, 22,* 483–510.

Ialongo, N., Poduska, J., Werthamer, L., & Kellam, S. (2001). The distal impact of two first-grade preventive interventions on conduct problems and disorder in early adolescence. *Journal of Emotional & Behavioral Disorders, 9,* 146–161.

Ialongo, N. S., Rogosch, F. A., Cicchetti, D., Toth, S., Buckley, J. A., Petras, H., et al. (2006). A developmental psychopathology approach to the prevention of mental health disorders. In D. Cicchetti & D. Cohen (Eds.), *Developmental psychopathology* (2nd ed., pp. 968–1018). New York: Wiley.

Ialongo, N., Werthamer, L., Brown, C. H., Kellam, S., & Wai, S. B. (1999). The proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression and antisocial behavior. *American Journal of Community Psychology, 27,* 599–642.

Institute of Education Sciences. (2007). *Fast facts.* Retrieved February 3, 2008, from http://nces.ed.gov/fastfacts/display.asp?id=16

Izzo, C. V., Connell, J. P., Gambone, M. A., & Bradshaw, C. P. (2004). Understanding and improving youth development initiatives through evaluation. In S. F. Hamilton & M. A. Hamilton (Eds.), *Youth development handbook: Coming of age in American communities* (pp. 301–326). Thousand Oaks, CA: Sage.

Jessor, R., Turbin, M. S., & Costa, F. M. (1998). Risk and protection in successful outcomes among disadvantaged adolescents. *Applied Developmental Science, 2,* 194–208.

Jo, B., & Muthén, B. O. (2002). Longitudinal studies with intervention and noncompliance: Estimation of causal effects in growth mixture modeling. In N. Duan & S. Reise (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 112–139). Mahwah, NJ: Erlbaum.

Karweit, N., & Bond, M. A. (1993). *Manual for the story telling and retelling program.* Baltimore: Center for Research on Elementary and Middle Schools, Johns Hopkins University.

Kaufman, A. S., & Kaufman, N. L. (1985). *Kaufman Test of Educational Achievement.* Circle Pines, MN: American Guidance Service.

Kellam, S. G., & Anthony, J. (1998). Targeting early antecedents to prevent tobacco smoking: Findings from an epidemiologically-based randomized field trial. *American Journal of Public Health, 88,* 1490–1495.

Kellam, S. G., Branch, J. D., Agrawal, K. C., & Ensminger, M. E. (1975). *Mental heath and going to school: The Woodlawn program of assessment, early intervention, and evaluation.* Chicago: University of Chicago Press.

Kellam, S. G., Brown, C. H., Poduska, J. M., Ialongo, N. S., Wang, W., Toyinbo, P., et al. (2008). Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. *Drug and Alcohol Dependence, 95*(Suppl. 1), 5–28.

Kellam, S. G., Mayer, L. S., Rebok, G. W., & Hawkins, W. E. (1998). The effects of improving achievement on aggressive behavior and of improving aggressive behavior on achievement through two prevention interventions: An investigation of causal paths. In B. Dohrenwend (Ed.), *Adversity, stress, and psychopathology* (pp. 486–505). New York: Oxford University Press.

Kellam, S. G., & Rebok, G. W. (1992). Building developmental and

etiological theory through epidemiologically based preventive intervention trials. In J. McCord & R. E. Tremblay (Eds.), *Preventing antisocial behavior: Interventions from birth through adolescence* (pp. 162–195). New York: Guilford Press.

Kellam, S. G., Rebok, G. W., Ialongo, N., & Mayer, L. S. (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry, 35,* 259–281.

Kistner, J. A., David, C. F., & White, B. A. (2003). Ethnic and sex differences in children's depressive symptoms: Mediating effects of perceived and actual competence. *Journal of Clinical Child and Adolescent Psychology, 32,* 341–350.

Lipsey, M. W. (1998). Design sensitivity: Statistical power for applied experimental research. In L. Bickman & D. L. Rog (Eds.), *Handbook of applied social research methods* (pp. 39–68). Thousand Oaks, CA: Sage.

McIntosh, K., Horner, R. H., Chard, D. J., Boland, J. B., & Good, R. H. (2006). The use of reading and behavior screening measures to predict nonresponse to school-wide positive behavior support: A longitudinal analysis. *School Psychology Review, 35,* 275–291.

McKay, M. M., Hibbert, R., Hoagwood, K., Rodriguez, J., Murray, L., Legerski, J., et al. (2004). Integrating evidence-based engagement interventions into "real world" child mental health settings. *Brief Treatment and Crisis Intervention, 4,* 177–186.

Miller-Cribbs, J. E., Cronen, S., & Davis, L. (2002). An exploratory analysis of factors that foster school engagement and completion among African American students. *Children & Schools, 24,* 159–174.

Mrazek, P. G., & Haggerty, R. J. (Eds.). (1994). *Reducing risks for mental disorders: Frontiers for preventive intervention research.* Washington, DC: National Academy Press.

Osterholm, K., Nash, W. R., & Kritsonis, W. A. (2007). Effects of labeling students "Learning Disabled": Emergent themes in the research literature 1970 through 2000. *Focus on Colleges, Universities, and Schools, 1,* 1–11.

Patrikakou, E. N., & Weissberg, R. P. (2007). School–family partnerships and children's social, emotional, and academic learning. In R. Bar-on, J. G. Maree, & M. J. Elias (Eds.), *Educating people to be emotionally intelligent* (pp. 55–67). Rondebosch, South Africa: Heinemann Educational Publishers.

Patterson, G. R., DeBaryshe, B. D., & Ramsey, E. (1989). A developmental perspective on antisocial behavior. *American Psychologist, 44,* 329–335.

Patterson, G. R., Reid, J., & Dishion, T. (1992). *Antisocial boys.* Eugene, OR: Castalia.

Perie, M., Grigg, W. S., & Donahue, P. L. (2005). *Nation's report card: Reading 2005* (NCES Rep. No. 2006–451). Washington, DC: U.S. Department of Education Institute of Education Sciences.

Petras, H., Chilcoat, H., Leaf, P., Ialongo, N., & Kellam, S. (2004). The utility of teacher ratings of aggression during the elementary school years in identifying later violence in adolescent males. *Journal of the American Academy of Child and Adolescent Psychiatry, 1,* 88–96.

Prevatt, F., & Kelly, F. D. (2003). Dropping out of school: A review of intervention programs. *Journal of School Psychology, 41,* 377–395.

Reid, J., Eddy, M., & Fetrow, R. (1999). Description and immediate impacts of a preventive intervention for conduct problems. *American Journal of Community Psychology, 27,* 483–517.

Schaeffer, C. M., Petras, H., Ialongo, N., Masyn, K. E., Hubbard, S., Poduska, J., et al. (2006). A comparison of girls' and boys' aggressive–disruptive behavior trajectories across elementary school: Prediction to young adult antisocial outcomes. *Journal of Consulting and Clinical Psychology, 74,* 500–510.

Schaeffer, C., Petras, H., Ialongo, N., Poduska, J., & Kellam, S. (2003). Modeling growth in boys' aggressive behavior across elementary school: Links to later criminal involvement, conduct disorder, and antisocial personality disorder. *Developmental Psychology, 39,* 1020–1035.

Schwartz, D., Gorman, A. H., Duong, M. T., & Nakamoto, J. (2008). Peer relationships and academic achievement as interacting predictors of depressive symptoms during middle childhood. *Journal of Abnormal Psychology, 117,* 289–299.

Scott, T., & Barrett, S. (2004). Using staff and student time engaged in disciplinary procedures to evaluate the impact of school wide PBS. *Journal of Positive Behavior Interventions, 6,* 21–27.

Scott-Jones, D. (1984). Family influences on cognitive development and school achievement. *Review of Research in Education, 11,* 259–304.

Smart, D., Sanson, A., & Prior, M. (1996). Connections between reading disability and behavioral problems: Testing temporal and causal hypotheses. *Journal of Abnormal Child Psychology, 24,* 363–375.

Stenmark, J. K., Thompson, V., & Cossey, R. (1986). *Family math.* Berkley: University of California.

Storr, C., Ialongo, N., Anthony, J., & Kellam, S. (2002). A randomized prevention trial of early onset tobacco use by school and family-based interventions implemented in primary school. *Drug and Alcohol Dependence, 66,* 51–60.

Sugai, G., & Horner, R. (2006). A promising approach for expanding and sustaining school-wide positive behavior support. *School Psychology Review, 35,* 245–259.

Vernez, G., Krop, R. A., & Rydell, C. P. (1999). *Closing the education gap: Benefits and costs.* New York: Rand.

Walker, H. M., Ramsay, E., & Gresham, F. M. (2004). *Antisocial behavior in school: Evidence-based practices* (2nd ed.). Belmont, CA: Wadsworth/Thomson Learning.

Webster-Stratton, C. (1984). Randomized trial of two parent-training programs for families with conduct-disordered children. *Journal of Consulting and Clinical Psychology, 52,* 666–678.

Werthamer-Larsson, L., Kellam, S. G., & Wheeler, L. (1991). Effect of first-grade classroom environment on child shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology, 19,* 585–602.

Whitehurst, G. J., Epstein, A. L., Payne, A. C., Crone, D. A., & Fischel, J. E. (1994). Outcomes of an emergent literacy intervention in Head Start. *Journal of Educational Psychology, 86,* 542–555.

Worthington, C. F. (1987). Taking the test: Kaufman Test of Educational Achievement, Comprehensive Form and Brief Form. *Journal of Counseling and Development, 65,* 325–327.

Received July 29, 2008
Revision received May 4, 2009
Accepted May 8, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 938–947

0022-0663/09/$12.00 DOI: 10.1037/a0016212

Syllable and Letter Knowledge in Early Korean Hangul Reading

Jeung-Ryeul Cho
Kyungnam University

This study examined the development of Korean consonant–vowel (CV) syllable identification, consonant and vowel letter knowledge, and their relationships to phonological awareness and the reading of regular Hangul words among Korean kindergartners as a 6-month longitudinal study. Results showed that Korean children identified CV syllables better than consonant and vowel letters. In regression analyses, CV syllable identification at Time 1 strongly contributed to Hangul word recognition concurrently over and above letter knowledge, as well as longitudinally after controlling for letter knowledge and reading at Time 1. However, letter knowledge did not predict Hangul reading once CV syllable identification was controlled. In addition, CV syllable knowledge facilitated subsequent letter knowledge and phoneme onset and coda awareness. The results, in general, shed light on the salient roles of syllables in the early literacy development of Korean.

Keywords: syllable, consonant- and vowel-letter knowledge, Korean Hangul reading

Supplemental materials: http://dx.doi.org/10.1037/a0016212.supp

The knowledge of letters during the kindergarten years has been known as one of the best longitudinal predictors of future reading ability in English as well as in other alphabetic languages (Adams, 1990; McBride-Chang, 2004). However, the current literature has not yet found similar evidence for languages other than Indo-European. Because reading processes may differ substantially depending on the writing systems, learning environment, and teaching methods (e.g., Harris & Hatano, 1999; Joshi & Aaron, 2006), the current research is based on the premise that it would be interesting to examine the role of letter knowledge in the early reading of non-European languages, such as Korean.

More specifically, as an alphabetic orthography, Korean Hangul uses 19 consonant letters and 21 vowel letters, with each representing a phoneme. Although Hangul is an alphabet, syllables are also important phonological and orthographic units in Korean (Cho & McBride-Chang, 2005a; Simpson & Kang, 2004). In particular, consonant–vowel (CV) open syllables are considered to be of a salient grain size of reading (Yi, 1998; Yoon, Bolger, Kwon, & Perfetti, 2002; Ziegler & Goswami, 2005). Thus, Korean is often called an *alphabetic-syllabary* (Taylor & Taylor, 1995). The current 6-month short-term longitudinal study explored the extent to which CV syllable and alphabet letter knowledge might predict Hangul word recognition among young Korean children. In addition, the study examined the effects of CV syllable and letter knowledge on subsequent letter knowledge and phonological awareness.

This study was supported by the 2006 Kyungnam University Foundation Grant. I thank Soon-Gil Park, Dal-Lae Jin, Young-Mee Paek, In-Suck Whang, and Jee-Hyun Lee for their assistance with the data and the students of Hansarang, Yoosung, and Daeja Kindergartens in Korea for their participation. I very much appreciate the suggestions from Catherine McBride-Chang and Roman Taraban on drafts of this article.

Correspondence concerning this article should be addressed to Jeung-Ryeul Cho, Division of Psychology and Sociology, Kyungnam University, Masan 631-701, South Korea. E-mail: jrcho@kyungnam.ac.kr

Korean Hangul Alphabet and Reading Acquisition

Korean Hangul consists of 14 basic and 5 doubled consonant letters, as well as 10 basic and 11 compound vowel letters.[1] Korean consonants were created in a unique way such that the 5 simple consonant letters (ㄱ, ㄴ, ㅁ, ㅅ, ㅇ) indicated the shapes of the articulators and places of articulation in pronouncing the sounds they stand for (J. Kim & Davis, 2006; Taylor & Taylor, 1995). For example, the letter ㄴ/n/ illustrates a tongue tip touching the upper gum, whereas the letter ㅁ/m/ depicts a closed mouth. The simple consonant letters were used to derive complex letters standing for aspiration and tenseness by adding strokes.[2] For example, the aspirated consonant ㅋ/k/ is created by adding one stroke inside the letter ㄱ/g/. Individual consonant and vowel letters are visually distinctive and composed of one to four strokes with vertical and horizontal lines and/or a circle. Although fundamentally alphabetic, Hangul letters are grouped into syllable blocks. Each block is separated by a small space; a larger space divides words. The syllable block printing of Hangul makes syllables visually identifiable. For example, "Hangul" in Korean (한글) is written in two-syllable blocks, instead of a linear arrangement (ㅎㅏㄴㄱㅡㄹ) as in English. As shown in the example, the reading order of the letters within a syllable block is systematic and is from top to bottom and from left to right. The sound structure of Korean is fairly simple, with either a CV or a CVC structure.

Unlike English, Korean Hangul is classified as a shallow orthography, with more regular grapheme-to-phoneme correspondences. In English, letter names are mostly monosyllabic. Most English letter names contain corresponding letter sounds, either as an initial phoneme (CV letter names, e.g., 'b,' 'd,' 'p') or as a final

[1] Hangul consonants, vowels, and syllables are available in the supplemental materials.

[2] The principles of adding strokes in the Korean alphabet are provided in the supplemental materials.

phoneme (VC letter names, e.g., 'f,' 'l,' 'm'). Initial sounds in CV letter names are learned earlier than are final sounds (Treiman, Tincoff, & Richmond-Welty, 1997; Treiman, Weatherston, & Berch, 1994). In contrast, letter names in the Korean language are complicated, as in Hebrew (e.g., Levin, Patel, Margalit, & Barad, 2002). Each Korean consonant has a name and represents one sound, whereas each vowel has the same name as the sound it represents. The names of Korean basic consonants consist of two syllables mostly in a CV–VC form. For instance, the letter 'ㄴ' has the name 니은/ni.ɨn/ and represents the sound /n/, whereas the letter 'ㄹ' has the name 리을/li.ɨl/ with the sound /l/. In general, each consonant name starts with its own sound and ends with the sound value of the letter made at the syllable-final position. Similarly, the names of Hebrew letters are based on the acrophonic principle, which tends to make the connection between each letter name and sound clear. It also facilitates learning the alphabetic principle (Levin et al., 2002; Share & Levin, 1999). For this reason, consonant names in Korean may be especially helpful in acquiring letter sounds as well as in reading acquisition.

The presence of complex consonant names as well as a relatively simple syllable structure in Korean may affect how children learn to read Hangul. Unlike in other alphabetic orthographies in which children first learn the alphabet letters in their own names, Korean children first learn to read the printed syllables in their names. For example, Yoon (1997) found that 3-year-old Korean children read 61% of syllables included in their own names and 46% of CV syllables; 4-year olds read 91% of CV syllables and 41% of CVC syllables; 5-year olds read 72% of CVC syllables. Moreover, in another study that investigated Korean children's letter knowledge, 4-year olds identified 26% of the basic consonant names and 14% of the basic consonant and vowel sounds, whereas 5-year-old children identified 61% and 58% of basic letter names and sounds, respectively (Choi & Yi, 2007). When they enter a primary school, at 6 years of age most children master reading regular words (e.g., Cho & McBride-Chang, 2005a) and reading one-syllable pseudowords (Yoon, 1997). In primary schools, children are taught various morphophonological rules that alter the phonology of some words, such as through resyllabification, simplications of multiple coda, and consonantal assimilation.

Although Korean children first read syllables earlier than words and alphabet letters, the teaching methods of Korean Hangul are diverse. A whole-word method has been adopted as a Korean government policy for Hangul literacy education since the 1990s. Korean kindergartens are not supposed to teach decoding skills, but they enhance children's interest and motivation to read and write (Ministry of Education and Human Resource Development, 1998). Kindergartens use the whole-word approach in which teachers present children with stories to read and instruct them to read and spell words. In addition, about 85% of children have been given extra literacy training at home, such as studying commercially available workbooks themselves as home education or enrolling in commercial institutions (e.g., Korean Association of Child Studies & Hansol Education Research Center, 2002). These types of after-school reading programs often employ phonics instruction (Lee & Lee, 2007). Although a CV syllable chart is not popularly used these days,[3] it was favored as an effective method for less educated people before the 20th century, and its use was re-emphasized again in the 1960s (e.g., Taylor & Taylor, 1995). In a typical CV chart, 14 basic consonant letters are arranged in columns and 10 basic vowel letters are arranged in rows to form 140 CV syllable blocks. With the CV chart, children are often taught to read all of the syllable blocks in the order of the consonants (e.g., 가나다라마바사아자차카타파하) or vowels (가갸거겨고교구규그기). CV syllables are always regular in their pronunciation in that they follow the grapheme-to-phoneme correspondence rules of Korean, whereas the reading of multisyllable words is often subject to phonological changes as a result of the morphophonemic writing and assimilation phenomena of the Korean language.

A substantial amount of research on reading acquisition has consistently found letter-name knowledge to be a powerful preschool predictor of reading and initial spelling achievement (e.g., Levin et al., 2002; Muter, Hulme, Snowling, & Taylor, 1997; Pennington & Leftly, 2001; Shatil, Share, & Levin, 2000). For example, Share, Jorm, Maclean, and Matthews (1984) found that letter-name knowledge in kindergarten was the best individual predictor of kindergarten reading achievement among 39 variables, including IQ, vocabulary, and home environment, and the second best predictor of first-grade reading achievement after phoneme segmentation. Evidence also indicates that letter-name knowledge precedes and facilitates letter-sound knowledge (see Foulin, 2005, for a review; Share, 2004; Treiman et al., 1994). When a letter name includes letter sounds, as in many alphabetic languages, letter-name knowledge often serves as background knowledge that children rely on to easily acquire letter–phoneme correspondences. Such assistance of letter-name knowledge is stronger when the initial phoneme of the letter's name corresponds to the letter's sound (Levin et al., 2002; Treiman, Tincoff, & Richmond-Welty, 1996). Letter sound knowledge is also widely recognized as necessary for acquiring the alphabetic principle and learning to read alphabetic texts successfully (Stuart & Coltheart, 1988; Treiman et al., 1994). Along similar lines, Korean alphabet knowledge was also found to be associated with Hangul reading among kindergartners (Choi & Yi, 2007).

In addition to letter knowledge, phonological awareness was found to be an important factor in the development of literacy skills in English as well as in Korean (Adams, 1990; Burgess & Lonigan, 1998; Cho & McBride-Chang, 2005a; Ehri, Nunes, & Willows, 2001; McBride-Chang, 2004; Wagner, Torgesen, & Rashotte, 1994). Phonological awareness refers to the ability to detect and manipulate the sound units of one's oral language ranging from syllables to phonemes. For instance, Hangul word recognition among young children was associated with both phoneme and syllable awareness in a cross-sectional study (Cho & McBride-Chang, 2005a) and with only syllable awareness in a 1-year longitudinal study (Cho & McBride-Chang, 2005b).

Recently, Cho, McBride-Chang, and Park (2008) demonstrated that phonological awareness at the three levels of syllable, onset,

[3] Because there has been no literature on the method of using CV syllable charts in kindergartens, a brief survey was conducted with 29 kindergarten teachers in a city in Korea. As results, 100% of the teachers responded that they presented children with stories to read, 81% of them taught the children to read and spell words, 48% taught syllable reading and spelling, 44% taught alphabet letters, and 31% used CV syllable charts.

and coda was significantly associated with the reading of phonologically regular Hangul words, whereas other skills, such as morphological awareness and visual skills, were related to reading irregular words. This study, in particular, focused on the reading of regular Hangul words that young Korean children most often encounter at an early stage of Hangul reading acquisition. According to dual-route models, the recognition of Hangul words, especially regular words, is assumed to rely on phonologically mediated sublexical pathways, as Korean is a relatively shallow orthography (e.g., Cho & Chen, 1999; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Seidenberg, 1985). In these models, phonological information is assembled in sublexical pathways on the basis of grapheme–phoneme correspondences or some other information, whereas it is retrieved from a lexicon stored in memory in the lexical pathways.

In particular, syllabic information was emphasized as an important sublexical unit in Korean Hangul word recognition. For example, Simpson and Kang (2004) demonstrated the strong effects of syllable frequencies as compared with the minimal effects of word and subsyllabic frequencies on naming Hangul stimuli among adults. This indicates that the Korean syllables serve as functional units in Hangul word recognition. The syllable body (onset plus vowel) was found to be prominent in both the reading and speaking of Korean (Yi, 1998; Yoon et al., 2002), whereas rime (vowel plus coda) is important in many other languages, such as English, German, Dutch, and French (e.g., Taraban & McClelland, 1987). For example, as found in Yoon et al. (2002), 4- and 5-year-old Korean children could read nonwords two times better when clue words and test nonwords shared the CV body than when they shared the rime. The Korean children in their study performed similarly in a grapheme substitution task using English with linear Roman letters. Their results suggest that the salience of CV units is independent of the form of printed Hangul and is instead driven by the intrinsic features of the Korean language, such as the high frequency of CV syllable types (e.g., Y.-S. Kim, 2007). In this regard, the CV syllable is suggested as a salient grain size in Korean (Ziegler & Goswami, 2005). The present study further examined the extent to which CV syllable and alphabetic knowledge in Korean contributes to the reading of Hangul words among Korean early readers.

CV syllables and letter knowledge are likely to influence the development of phonological awareness. According to the studies of English and other alphabetic languages, earlier letter knowledge contributed to the development of phoneme awareness, and these two skills were closely interrelated during the preschool years (Burgess & Lonigan, 1998; Foulin, 2005; Lonigan, Burgess, & Anthony, 2000; Wagner et al., 1994, 1997). Several studies further indicate that letter-sound knowledge was more strongly associated with phoneme awareness than was letter-name knowledge (McBride-Chang, 1999; Read, Zhang, Nie, & Ding, 1986). This study thus investigated how well CV syllable and letter knowledge is longitudinally predictive of letter knowledge and phonological awareness in Korean.

The present study was a 6-month short-term longitudinal study: Korean kindergartners were initially tested at the end of the kindergarten year, and a follow-up test was conducted 6 months later. CV syllables, consonant and vowel knowledge, and Hangul word reading were assessed at Times 1 and 2, phonological awareness at Time 2 only, and oral vocabulary at Time 1 only. The study explored three specific issues. First, how much do young Korean children learn about CV syllables and consonant and vowel letters? To the best of my knowledge, this is the first study that examines both syllable and alphabetic knowledge in Korean, as related studies in the literature examined either syllable reading or letter knowledge separately. On the basis of the salience of CV subsyllabic body in the Korean language (Yi, 1998; Yoon et al., 2002), I predicted that Korean children would identify CV syllables better than letter names and sounds. The second research question explored was, To what extent would CV syllables and letter knowledge predict the concurrent and subsequent reading of regular Hangul words? If regular Hangul words are processed through sublexical syllabic information (e.g., Simpson & Kang, 2004) and CV syllables are basic orthographic units (e.g., Yoon et al., 2002; Ziegler & Goswami, 2005), then knowledge of CV syllables would be expected to predict the reading of regular Hangul words. If this is the case, the extent to which letter knowledge contributes to the reading of Hangul words would be relatively limited. The third question addressed by this study was the following: How do CV syllables and letter knowledge predict subsequent letter knowledge and phonological awareness? Knowledge of the Korean letter sounds have been demonstrated to be associated with phonological awareness (e.g., Burgess & Lonigan, 1998; McBride-Chang, 1999; Wagner et al., 1994). However, the extent to which early CV syllable identification predicts subsequent letter knowledge and phonological awareness has been unclear.

Method

Participants

Participants were 189 kindergartners (101 boys and 88 girls), with a mean age of 5.98 years (SD = 0.61) at Time 1. The participants attended classes for 4-year-olds (kindergarten year 1, hereinafter referred to as K1; N = 85, mean age = 5.37 years) and 5-year olds (kindergarten year 2, hereinafter referred to as K2; N = 104, mean age = 6.43 years). All participants were Korean native speakers from two private kindergartens for 4-year-olds and one public kindergarten for 5-year-olds in Masan, a medium-sized city in South Korea. Most children were from middle-class backgrounds. Note that children from kindergartens located in cities tend to be from middle-class families, whereas many children of low-income families and working mothers are normally enrolled in welfare and childcare facilities (referred to as Children's House) because of their full-day operation (Na & Moon, 2004). The children were initially tested in January 2005 and followed up 6 months later. At Time 2, K1 students moved on to classes of 5-year-olds in the same kindergarten, whereas K2 students entered public primary schools and were given 5 months of instruction. Because of dropouts at Time 2 testing, 166 children (for K1, N = 71; for K2, N = 95) were tested at Time 2. No significant differences were found between the participants included at Time 2 and the dropouts in terms of age and scores at Time 1 ($ps > .05$). The three kindergartens used the whole-word approach to teaching reading: As in most Korean kindergartens, teachers presented children with stories to read and taught them to read and spell words without explicitly teaching alphabet letters and without using a CV chart. In addition, over 80% of preschool children were reported to be given extra literacy training at home (e.g., Lee & Lee, 2007). Many of the participants in this study were also

learning letter names and sounds at home before entering primary school. However, the level of literacy training of the children at home was not measured in this study.

Procedure

The children were tested at school by a research team that consisted of two graduate students and four undergraduate students majoring in psychology. The psychology students were well trained on how to test the children and how to record the data. Each child individually participated in two testing sessions, each lasting approximately 1–1.5 hr (including several other tasks that went beyond the scope of this study but were part of a longitudinal study). All children were administered the tasks described below.

Measures

Hangul word reading. For both testing times, children were given the same test to read 35 two-syllable words. The level of difficulty gradually increased with the introduction of compound vowels (e.g., ㅑ/ja/, ㅙ/wæ/) and consonants (e.g., ㅄ, ㅉ). The items could be read correctly by applying the Korean grapheme-to-phoneme conversion rules. After children were asked to begin reading, experimenters stopped testing when five consecutive items were read incorrectly. Feedback was not given to the children. One point was allotted for each word correctly read aloud.

CV syllable identification. For both testing times, children were given the same test of 19 CV syllables. Each of all 19 consonants (14 basic and 5 doubled) was combined with 1 of 10 basic vowels, each of which was used once or twice, to construct the 19 CV syllables used in this study. The level of difficulty was increased by introducing basic consonants at the beginning of the list and doubled consonants toward the end. Testing was stopped when a participant had made errors on five consecutive items. One point was given for each syllable correctly read aloud.

Consonant- and vowel-letter knowledge. At both testing times, children were tested on three tasks: 19 consonant names, 19 consonant sounds, and 19 vowel sounds. The consonants consisted of 14 basic and 5 doubled consonants, and the vowels were divided into 10 basic and 9 compound vowels. Each task presented the basic letters prior to the compound or doubled letters, and the letters in each category of either basic or compound letters were presented in a random order. The first three items of each task were presented with feedback. In the testing of consonant sounds, most K1 children made errors in the first testing item because they did not understand the instruction to identify the sound made by each consonant letter. Thus, the first item of the consonant-sound task was excluded from further analyses. Consequently, the maximum score was 19 in the consonant-name and vowel-sound tasks and 18 in the consonant-sound task.

Phonological awareness. The three tasks of syllable, onset, and coda deletion were included. During the syllable deletion task, at Time 2 only, children were asked to listen to three-syllable words and nonwords that were orally presented. From each three-syllable stimulus item, children were asked to delete one syllable; for example, e.g., *mog so ri* (목소리) without *mog* (목) would be *so ri* (소리). The task was designed to present items increasing in difficulty by including 8 real words first and 8 nonwords toward the end, for a total of 16 items in this task. Of the 16 items of this task, 8 required the deletion of the middle syllable. Of the remaining 8 items, 4 items required removing the first syllable and another 4 required eliminating the last syllable. The maximum score was 16.

In the phoneme onset deletion task, at Time 2 only, children were asked to delete the first phoneme from one-syllable real words that were orally presented. For example, saying *gam* (감) without the initial sound would be *am* (암). Of the items administered, 8 were CV syllables, and the other 8 were CVC syllables. Items increased in difficulty level with the introduction of CV words first and CVC words toward the end of testing. The maximum score of this task was 16.

In the coda deletion task, at Time 2 only, children were asked to delete the final phoneme from one-syllable CVC real words that were orally presented. For example, saying *non* (논) without the final sound would be *no* (노). In total, 16 items were included in this task. Items increased in difficulty with the inclusion of diphthongs and compound consonants.

In these three tasks, experimenters stopped testing when five consecutive items were failed. Feedback was not given to the children.

Vocabulary. At Time 1 only, the Korean-Wechsler Preschool and Primary Scale of Intelligence Vocabulary subtest was administered to measure general vocabulary skills (Park, Kwak, & Park, 1995). In the test, children were asked to define or explain pictures of objects and more difficult concepts. Testing was stopped when a participant scored 0 marks for five consecutive items. The scoring procedure of this task was based on the local norm established by the authors to calculate standardized scores by age.

Results

The means and standard deviations are presented in Table 1. In general, the tasks showed adequate internal consistency reliabilities. Table 2 shows how Time 1 CV syllable identification and letter knowledge were correlated with Time 2 variables and Time 1 reading task, when grade, age, and vocabulary were statistically controlled. CV syllable identification at Time 1 was strongly correlated with reading at Times 1 and 2 (rs = .82 and .77, $ps < .001$), and Time 1 letter knowledge was moderately associated with reading at both times ($.18 < rs < .43$, $ps < .05$). Time 1 CV syllable identification was significantly correlated with Time 2 letter knowledge ($.39 < rs < .46$, $ps < .001$) and phonological awareness ($.22 < rs < .41$, $ps < .01$). Time 1 consonant naming was moderately correlated with consonant- and vowel-sound knowledge at Time 2 (r = .32 in both, $ps < .001$).

Performance on Consonant–Vowel Syllables and Letter Knowledge

Table 3 shows the percentages of K1 and K2 children who correctly identified CV syllables, consonant names, consonant sounds, and vowel sounds across Times 1 and 2. Comparisons in performance show that children generally identified CV syllables better than consonant names ($ts > 2.45$, $ps < .05$), except for K1 children at Time 2. In addition, consonant-name identification was better than consonant-sound identification in all comparisons ($ts > 2.25$, $ps < .05$). The difference between consonant name and

Table 1
Reliabilities, Means, and Standard Deviations for the Variables Tested at Times 1 and 2

Variable	*M*	*SD*	Reliability
Age (months)	71.73	7.27	
Word reading (35)			
Time 1	26.24	10.62	.98
Time 2	31.45	6.99	.97
Vocabulary (18)			
Time 1	13.16	2.77	.59
Consonant–vowel syllable identification (19)			
Time 1	16.82	4.06	.94
Time 2	17.35	3.15	.92
Consonant name (19)			
Time 1	12.63	5.63	.92
Time 2	16.87	3.93	.93
Consonant sound (18)			
Time 1	9.10	7.63	.98
Time 2	14.10	6.22	.98
Vowel sound (19)			
Time 1	10.69	5.97	.94
Time 2	14.63	4.64	.92
Syllable deletion (16)			
Time 2	10.16	4.44	.89
Onset deletion (16)			
Time 2	7.40	5.43	.95
Coda deletion (16)			
Time 2	12.55	5.31	.96

Note. $N = 189$ for Time 1 measures; $N = 166$ for Time 2 measures. The maximum score for each measure is indicated within parentheses.

consonant sound was moderate for K1 children, but such difference was smaller for K2 children. Vowel-sound knowledge was better than consonant-sound knowledge for K1 children at Time 1 only, $t(84) = 3.02$, $p < .01$. Thus, acquiring sound knowledge of Korean consonants and vowels appears to be a challenge relative to consonant naming for Korean children. The performance on CV syllables was better than that on the three tasks of letter knowledge for both grades.

Table 3 also presents the percentages of K1 and K2 children who correctly identified basic, aspirated, and tense consonants and basic and compound vowels. From the table, I compared the performance for the different letter categories, which were grouped on the basis of stroke-adding principles of the Korean alphabet (e.g., Taylor & Taylor, 1995). The performance of basic consonants was better than that of the aspirated and tense consonants for the two grades at both testing times. K2 children had learned more than 90% of the basic consonant and vowel letters at Time 2 by the time they entered primary school, but they had not yet mastered the compound consonants and vowels. Specifically, learning doubled and tense consonants was a challenge to the children. In addition, the correlations of consonant names and sounds with the alphabet order were significant ($rs > .54$, $ps < .05$) in all testing except for K2 children at Time 2. However, the correlations between vowel sounds and alphabet order were not significant for either grade. These results indicated that young children tended to learn consonant names and sounds from the beginning of the alphabet. However, this tendency was not supported in the identification of vowels.

Table 3
Percentages of Kindergarten Year 1 (K1) and Year 2 (K2) Children Who Correctly Identified Consonant–Vowel Syllables, Consonant and Vowel Groups at Times 1 and 2

	K1 children		K2 children	
Consonant–vowel syllable	Time 1	Time 2	Time 1	Time 2
Consonant–vowel syllable	78	84	96	98
Consonant name				
Basic (9)	65	91	84	99
Aspirated (5)	53	83	71	94
Tense (5)	31	74	68	84
All (19)	54	81	76	94
Consonant sound				
Basic (8)	35	76	74	93
Aspirated (5)	27	65	67	86
Tense (5)	21	60	57	78
All (18)	29	68	68	87
Vowel sound				
Basic (10)	53	85	85	98
Compound (9)	25	46	53	72
All (19)	39	66	70	85

Note. $N = 85$ in K1 at Time 1; $N = 71$ in K1 at Time 2; $N = 104$ in K2 at Time 1; $N = 95$ in K2 at Time 2. The number of items for each variable is indicated within parentheses.

Predicting Concurrent and Subsequent Reading of Hangul Words

The results of the analyses predicting concurrent and subsequent reading for the combined sample of children are presented in

Table 2
Correlations of Time 1 Variables With Time 2 Variables and Time 1 Reading Tasks, Controlling for Age, Grade, and Vocabulary

	Word reading		Consonant–vowel syllable	Consonant name	Consonant sound	Vowel sound	Syllable deletion	Onset deletion	Coda deletion
Variable	Time 1	Time 2	Time 2	Time 2	Time 2	Time 2	Time 2	Time 2	Time 2
Consonant–vowel syllable, Time 1	.82***	.77***	.67***	.46***	.39***	.43***	.22**	.31***	.41***
Consonant name, Time 1	.43***	.40***	.37***	.47***	.32***	.32***	.18*	.16*	.17*
Consonant sound, Time 1	.26***	.18*	.21**	.23**	.37***	.32***	.32***	.26**	.16*
Vowel sound, Time 1	.43***	.36***	.35***	.28***	.38***	.50***	.28***	.40***	.31***

* $p < .05$. ** $p < .01$. *** $p < .001$.

Tables 4 and 5, respectively.[4] As an examination of the extent to which CV syllables and letter knowledge at Time 1 contributed to Hangul reading at Time 1, hierarchical regression analyses were conducted by entering grade and age in Step 1 and vocabulary in Step 2 to control for their effects. When CV syllable identification scores were entered in Step 3 before entering the consonant-name, consonant-sound, and vowel-sound identification scores in Step 4, CV syllable identification explained 49% ($p < .001$) of the total variance over and above grade, age, and vocabulary, but letter knowledge did not account for a significant amount of variance after additionally controlling for CV syllable identification. When letter knowledge and CV syllable identification were entered in a different order in Steps 3 and 4, respectively, letter knowledge explained 16% ($p < .001$) of the total variance, and CV syllable identification accounted for an additional 33% ($p < .001$) of the variance after taking into consideration grade, age, vocabulary, and letter knowledge. Table 4 also summarizes the final standardized beta weights for each predictor included in the regression. The final beta weights revealed that only CV syllable identification uniquely contributed to concurrent reading after controlling for all other predictors.

In predicting subsequent reading at Time 2 in hierarchical regression analyses, grade and age were entered in Step 1, vocabulary in Step 2, and Time 1 reading in Step 3 to control for their effects. When CV syllable identification was entered in Step 4 and letter knowledge was entered in Step 5, CV syllable identification explained 7% ($p < .001$) of the total variance, but letter knowledge did not explain a significant amount of variance. When CV syllable identification and letter knowledge were included in a different order, letter knowledge did not explain a significant amount of variance, but CV syllable identification in Step 5 explained 7% ($p < .001$) of the total variance over and above all of the other predictors. Final beta weights showed that word recognition was predicted by Time 1 reading and CV syllable identification longitudinally.

Table 4
Hierarchical Regression Analyses Predicting Concurrent Reading at Time 1 From Time 1 Predictor Variables

Step/Variable	R^2 change	R^2	B	t
Step 1				
Grade	.16***	.16	.02	0.26
Age			−.05	−0.53
Step 2				
Vocabulary, Time 1	.00	.16	−.09	−1.38
Step 3				
Consonant–vowel syllable, Time 1	.49***	.65	.74	12.38***
Step 4				
Consonant name, Time 1	.01	.66	.07	1.04
Consonant sound, Time 1			.03	0.45
Vowel sound, Time 1			.01	0.13
Step 3				
Consonant name, Time 1	.16***	.33		
Consonant sound, Time 1				
Vowel sound, Time 1				
Step 4				
Consonant–vowel syllable, Time 1	.33***	.66		

Note. $N = 189$ for Time 1 measures.
*** $p < .001$.

Table 5
Hierarchical Regression Analyses Predicting Subsequent Reading at Time 2 From Time 1 Predictor Variables

Variable	R^2 change	R^2	B	t
Step 1				
Grade	.16***	.16	.01	0.07
Age			−.07	−0.81
Step 2				
Vocabulary, Time 1	.00	.16	−.11	−1.72
Step 3				
Reading, Time 1	.44***	.60	.32	3.36**
Step 4				
Consonant–vowel syllable, Time 1	.07***	.68	.51	5.74***
Step 5				
Consonant name, Time 1	.00	.68	.50	0.84
Consonant sound, Time 1			.01	0.12
Vowel sound, Time 1			−.01	−0.23
Step 4				
Consonant name, Time 1	.01	.61		
Consonant sound, Time 1				
Vowel sound, Time 1				
Step 5				
Consonant–vowel syllable, Time 1	.07***	.68		

Note. $N = 189$ for Time 1 measures; $N = 166$ for Time 2 measures.
** $p < .01$. *** $p < .001$.

Predicting Subsequent Letter Knowledge and Phonological Awareness

Table 6 presents the final standardized beta weights when variables tested at Time 1 were included in simultaneous regression equations. In the regression equations, CV syllable identification predicted a significant amount of variance in subsequent CV syllable identification. CV syllable and consonant-name identification independently predicted consonant naming at Time 2; CV syllables, consonant sounds, and vowel sounds uniquely predicted both consonant-sound and vowel-sound knowledge at Time 2. With all of the predictor variables included, the total amounts of variance explained in CV syllables, consonant names, consonant sounds, and vowel sounds at Time 2 were 57%, 39%, 35%, and 45%, respectively. Note that CV syllable identification predicted consonant and vowel knowledge longitudinally but that the prediction in the opposite direction was not significant.

Table 7 summarizes the final standardized beta weights for each Time 1 variable predicting syllable, onset, and coda deletion at

[4] Hierarchical regression analyses were carried out separately for K1 and K2 children. The main findings were the same in the two groups. For example, CV syllables explained significant variance in reading concurrently (for K1 children, 26%; for K2 children, 41%) after controlling for grade, age, vocabulary, and letter knowledge, as well as longitudinally (for K1 children, 5%; for K2 children, 6%) after additionally controlling for Time 1 reading. However, letter knowledge did not account for significant variance in reading after controlling for CV syllable identification and other predictors.

Table 6
Standardized Beta Weights for Regression Equations Predicting Consonant–Vowel Syllable Identification and Letter Knowledge at Time 2 From Time 1 Predictor Variables

	Time 2 variable							
	Consonant–vowel syllable		Consonant name		Consonant sound		Vowel sound	
Variable	ß	*t*	ß	*t*	ß	*t*	ß	*t*
Grade	.09	0.83	.15	1.20	.10	0.81	.00	0.02
Age	−.01	−0.10	−.06	−0.51	−.13	−1.05	−.01	−0.08
Vocabulary, Time 1	−.06	−0.86	.11	1.27	.00	−0.03	−.09	−1.08
Consonant–vowel syllable, Time 1	.61	9.16***	.33	4.16***	.25	3.00**	.27	3.49**
Consonant name, Time 1	.07	0.95	.34	4.04***	.06	0.75	−.00	−0.04
Consonant sound, Time 1	.05	0.83	.07	0.94	.27	3.39**	.15	2.04*
Vowel sound, Time 1	.05	0.68	−.03	−0.37	.19	2.01*	.37	4.30***

Note. $N = 189$ for Time 1 measures; $N = 166$ for Time 2 measures. $R^2 = .57$ for consonant–vowel syllable identification; $R^2 = .39$ for consonant-naming; $R^2 = .35$ for consonant sounds; $R^2 = .45$ for vowel sounds.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Time 2 in simultaneous regression equations. Consonant-sound knowledge accounted for a significant amount of variance in subsequent syllable deletion. CV syllables and vowel-sound knowledge predicted onset deletion at Time 2; vocabulary, CV syllables, and vowel-sound knowledge predicted coda deletion longitudinally. With all predictor variables included, the total amounts of variance explained in syllable, onset, and coda deletion at Time 2 were 26%, 35%, and 39%, respectively.

Discussion

The present study demonstrated the salient role of CV syllable identification on the recognition of regular Hangul words, both concurrently and longitudinally. However, letter knowledge did not predict Hangul reading once the CV syllable identification was controlled. CV syllable knowledge also facilitated growth in subsequent letter knowledge and phoneme onset and coda awareness. These findings are discussed in more detail next.

Korean children learned CV syllables earlier than letter knowledge, which supports early syllable reading by Korean children (e.g., Yoon, 1997). Consonant names were easier to learn than consonant sounds (e.g., Foulin, 2005; Share, 2004; Treiman et al., 1994). Additionally, consonant naming was moderately correlated with consonant sounds, suggesting that the knowledge of both overlap but are separate abilities (McBride-Chang, 1999; Treiman & Broderick, 1998; Treiman et al., 1994; Wagner et al., 1994). Korean consonant-letter knowledge and vowel-letter knowledge tend to develop in different ways. For example, both consonant-name and consonant-sound knowledge developed in the order of the Korean consonants, similar to English-speaking children's learning the English alphabet in order (e.g., McBride-Chang, 1999), whereas the learning of vowels was not influenced by the Korean vowel order. Children as young as 4 years of age in the first year of kindergarten knew more vowel sounds than consonant sounds, which may be due to the fact that Korean vowels have the same names as their sounds. However, this difference disappeared by 5 years of age. In addition, Korean children had mastered the basic consonant and vowel sounds in first grade, but they had not yet mastered the compound vowels and consonants, including

Table 7
Standardized Beta Weights for Regression Equations Predicting Syllable, Onset, and Coda Awareness at Time 2 From Time 1 Predictor Variables

	Time 2 variable					
	Syllable deletion		Onset deletion		Coda deletion	
Variable	ß	*t*	ß	*t*	ß	*t*
Grade	.08	0.58	.16	1.27	.20	1.64
Age	.04	0.30	−.02	−0.16	.14	1.16
Vocabulary, Time 1	.04	0.40	.02	0.22	.18	2.15*
Consonant–vowel syllable, Time 1	.12	1.30	.21	2.53*	.34	4.26***
Consonant name, Time 1	−.01	−0.12	−.13	−1.54	−.08	−1.00
Consonant sound, Time 1	.26	3.07**	−.15	1.82	.04	0.50
Vowel sound, Time 1	.17	1.68	.35	3.79***	.21	2.26*

Note. $N = 189$ for Time 1 measures; $N = 166$ for Time 2 measures. $R^2 = .26$ for syllable deletion; $R^2 = .35$ for onset deletion; $R^2 = .39$ for coda deletion.
* $p < .05$. ** $p < .01$. *** $p < .001$.

aspiration and tenseness. In particular, learning tense (i.e., doubled) consonants was a challenge to the children for several reasons, such as complexity in visual form and sound and the lack of explicit training of the stroke-adding principles of Korean in school (e.g., Taylor & Taylor, 1995).

Another important observation is that CV syllable and letter knowledge contributed differently to the reading of regular Hangul words. CV syllable identification accounted for a strong and unique amount of variance (33%) in the concurrent reading of Hangul words after statistically controlling for age, grade, vocabulary, and letter knowledge. More impressively, CV syllable identification explained a significant amount of variance (7%) in longitudinal reading after additionally controlling for Time 1 reading. However, letter knowledge did not predict a significant amount of variance in word recognition, either concurrently or longitudinally, after taking CV syllable identification into consideration. It is notable that almost 70% of the variance in concurrent and subsequent Hangul reading was accounted for cumulatively by the measures used in this study. This rate is higher than in previous studies of reading acquisition in Korean and English, with typically less than 50% of the variance in reading accounted for (e.g., Cho & McBride-Chang, 2005a; McBride-Chang, 1999; Torgesen et al., 1999). In particular, the limited role of Korean letter knowledge in Hangul word recognition is in sharp contrast to the previous findings from other alphabetic languages, in which letter knowledge has been considered one of the primary pre-reading abilities (e.g., Burgess & Lonigan, 1998; see Foulin, 2005, for a review; Levin et al., 2002; Muter et al., 1997; Share et al., 1984; Treiman et al., 1994). However, the strong and unique contribution of CV syllables in Hangul word recognition supports previous Korean studies and the psycholinguistic grain size theory (Simpson & Kang, 2004; Yoon et al., 2002; Ziegler & Goswami, 2005). Although Korean children prefer to use CV syllable knowledge in the recognition of regular Hangul words, CV syllables may not be a reliable cue in reading Korean exception words in which the consonant sounds of a syllable are often changed. Future research may pursue different decoding skills and strategies adopted by beginning readers in reading phonologically regular and irregular Hangul words (e.g., Cho et al., 2008).

In addition, CV syllable identification was found to be important in the subsequent development of letter knowledge, whereas letter knowledge was not predictive of subsequent CV syllable identification. Consonant naming did not contribute to consonant- and vowel-sound development in the current study (cf. Burgess & Lonigan, 1998; Foulin, 2005; McBride-Chang, 1999; Treiman et al., 1996). Few studies thus far have demonstrated that syllable identification is predictive of subsequent letter knowledge, although several studies have suggested that learning to read promotes letter knowledge and phonological sensitivity (Castles & Coltheart, 2004; Perfetti, Beck, Bell, & Hughes, 1987; Wagner et al., 1994). The strong contribution of CV syllables to letter knowledge, however, is not surprising if young Korean children have developed some implicit knowledge of letter names and sounds from learning CV syllables in which a consonant and a vowel are systematically combined. It would be interesting to explore the generality of these findings beyond the Korean language. For example, would similar results be found in Indo-European languages with linear letters? Such studies would shed light on how much of these effects can be attributed to characteristics of the Korean writing system in particular and how much can be accounted for by the salience of the syllables in general.

In the current study, letter-sound knowledge appeared to facilitate phonological awareness, which is consistent with previous English-language studies (e.g., McBride-Chang, 1999). Although the results are preliminary, Korean consonant- and vowel-sound knowledge was associated with different levels of phonological awareness: Consonant sounds were associated with syllable awareness, whereas vowel sounds were associated with phoneme onset and coda awareness. This indicates that syllable and phoneme awareness may be separate skills (e.g., Carroll, Snowling, Hulme, & Stevenson, 2003; Foy & Mann, 2001). Presumably, in the Korean language, a certain level of phoneme awareness skill requires the isolation of vowel sounds from syllables. Consonant- and vowel-letter knowledge may be bidirectionally related to phonological awareness in Korean language (e.g., Burgess & Lonigan, 1998). Their reciprocal relationship may be due to their associations with CV syllable identification in Korean (e.g., Perfetti et al., 1987; Wagner et al., 1994). However, this study only measured the effects of letter knowledge on phonological awareness, and did not include measurement of effects in the other direction, which was a limitation in the design of this study. As many studies have suggested that phonological awareness is a strong predictor of early reading skills, phonological awareness, particularly coda awareness, may facilitate later letter knowledge in Korean-language learning. The causal relationships of phonological awareness and letter-sound knowledge, along with CV syllables, should be studied in more detail in the future. In addition, the current study could not sort out the effect of teaching methods on the salience of CV syllables. The children's level of literacy training at home as well as their home background ideally should have been measured. Future studies may need to deliberately compare the reading behaviors of pre-readers or beginning readers taught with a whole-word method, a CV chart method, and a phonics method.

Despite these limitations, this study provides some practical implications for future educational and clinical work. The explicit teaching of children with CV syllables would probably be most effective for initial Korean Hangul reading acquisition. In particular, use of a CV syllable chart is recommended. Additionally, CV syllable training may be particularly useful in identifying Korean children who may have a potential problem in early language and literacy learning.

In conclusion, Korean children identified CV syllables earlier than consonant and vowel letters. CV syllable identification strongly and uniquely influenced the concurrent and subsequent reading of regular Hangul words, whereas the role of letter knowledge was found to be limited. In addition, CV syllable identification predicted subsequent consonant and vowel knowledge, as well as phoneme onset and coda awareness. The prominent role of CV syllables in early Hangul reading highlights the specific and unique features of Korean literacy development and raises interesting questions for further exploration of reading development in other languages.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Burgess, S. R., & Lonigan, C. J. (1998). Bidirectional relations of phono-

logical sensitivity and prereading abilities: Evidence from a preschool sample. *Journal of Experimental Child Psychology, 70,* 117–141.

Carroll, J. M., Snowling, M. J., Hulme, C., & Stevenson, J. (2003). The development of phonological awareness in preschool children. *Developmental Psychology, 39,* 913–923.

Castles, A., & Coltheart, M. (2004). Is there a causal link from phonological awareness to success in learning to read? *Cognition, 91,* 77–111.

Cho, J.-R., & Chen, H.-C. (1999). Orthographic and phonological activation in the semantic processing of Korean Hanja and Hangul. *Language and Cognitive Processes, 14,* 481–502.

Cho, J.-R., & McBride-Chang, C. (2005a). Correlates of Korean Hangul acquisition among kindergartners and second graders. *Scientific Studies of Reading, 9,* 3–16.

Cho, J.-R., & McBride-Chang, C. (2005b). Levels of phonological awareness in Korean and English: A 1-year longitudinal study. *Journal of Educational Psychology, 97,* 564–571.

Cho, J.-R., McBride-Chang, C., & Park, S.-G. (2008). Phonological awareness and morphological awareness: Differential associations to regular and irregular word recognition in early Korean Hangul readers. *Reading and Writing: An Interdisciplinary Journal, 21,* 255–274.

Choi, N.-Y., & Yi, S.-H. (2007). The effects of alphabet knowledge on Korean kindergartners' reading of Hangul words. *Journal of Korean Home Management Association, 25,* 151–168.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. C. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108,* 204–256.

Ehri, L. C., Nunes, S. R., & Willows, D. M. (2001). Phonological awareness instruction helps children learn to read: Evidence from the national Reading Panel's meta-analysis. *Reading Research Quarterly, 36,* 250–287.

Foulin, J. N. (2005). Why is letter-name knowledge such a good predictor of learning to read? *Reading and Writing, 18,* 129–155.

Foy, J. G., & Mann, V. (2001). Does strength of phonological representations predict phonological awareness in preschool children? *Applied Psycholinguistics, 22,* 301–325.

Harris, M., & Hatano, G. (Eds.). (1999). *Learning to read and write: A cross-linguistic perspective.* New York: Cambridge University Press.

Joshi, R. M., & Aaron, P. G. (Eds.). (2006). *Handbook of orthography and literacy.* Hillsdale, NJ: Erlbaum.

Kim, J., & Davis, C. (2006). Literacy acquisition in Korean Hangul: Investigating the perceptual and phonological processing of good and poor readers. In R. M. Joshi & P. G. Aaron (Eds.), *Handbook of orthography and literacy.* Hillsdale, NJ: Erlbaum.

Kim, Y.-S. (2007). Phonological awareness and literacy skills in Korea: An examination of the unique role of body-coda units. *Applied Psycholinguistics, 28,* 69–94.

Korean Association of Child Studies & Hansol Education Research Center. (2002). *Child development report 2001.* Seoul, Korea: Hansol Education.

Lee, K., & Lee, S. (2007). A study on the beginning literacy instruction through the analysis of commercial Korean language textbooks. *New Korean Education, 75,* 215–248.

Levin, I., Patel, S., Margalit, T., & Barad, N. (2002). Letter names: Effect on letter saying, spelling, and word recognition in Hebrew. *Applied Psycholinguistics, 23,* 269–300.

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study. *Developmental Psychology, 36,* 596–613.

McBride-Chang, C. (1999). The ABCs of the ABCs: The development of letter-name and letter-sound knowledge. *Merrill-Palmer Quarterly, 45,* 285–308.

McBride-Chang, C. (2004). *Children's literacy development.* London: Oxford University Press.

Ministry of Education and Human Resource Development. (1998). *Korean kindergarten curriculum.* Seoul, Korea: Special Education Publishing.

Muter, V., Hulme, C., Snowling, M., & Taylor, S. (1997). Segmentation, not rhyming, predicts early progress in learning to read. *Journal of Experimental Child Psychology, 65,* 370–396.

Na, J., & Moon, M. (2004). *Early childhood education and care policies in the republic of Korea* (OECD Thematic Review of Early Childhood Education and Care Policy: Background Report). Retrieved March 10, 2005, from http://www.oecd.org/dataoecd/25/57/27856763.pdf

Park, H. W., Kwak, K. J., & Park, K. B. (1995). *Korean-Wechsler Preschool and Primary Scale of Intelligence.* Seoul, Korea: Special Education Publishing.

Pennington, B. F., & Leftly, D. L. (2001). Early reading development in children at family risk for dyslexia. *Child Development, 72,* 816–833.

Perfetti, C. A., Beck, I., Bell, L., & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children. *Merrill-Palmer Quarterly, 33,* 283–319.

Read, C., Zhang, Y., Nie, H., & Ding, B. (1986). The ability to manipulate speech sounds depends on knowing alphabetic spelling. *Cognition, 24,* 31–44.

Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition, 19,* 1–30.

Share, D. L. (2004). Knowing letter names and learning letter sounds: A causal connection. *Journal of Experimental Child Psychology, 88,* 213–233.

Share, D. L., Jorm, A. F., Maclean, R., & Matthews, R. (1984). Sources of individual differences in reading acquisition. *Journal of Educational Psychology, 76,* 1309–1324.

Share, D. L., & Levin, I. (1999). Learning to read and write in Hebrew. In M. Harris & G. Hatano (Eds.), *Learning to read and write: A cross-linguistic perspective* (pp. 89–111). Cambridge, United Kingdom: Cambridge University Press.

Shatil, E., Share, D., & Levin, I. (2000). On the contribution of kindergarten writing to Grade 1 literacy: A longitudinal study in Hebrew. *Applied Psycholinguistics, 21,* 1–21.

Simpson, G. B., & Kang, H. (2004). Syllable processing in alphabetic Korean. *Reading and Writing: An Interdisciplinary Journal, 17,* 137–151.

Stuart, M., & Coltheart, M. (1988). Does reading develop in a sequence of stages? *Cognition, 30,* 139–181.

Taraban, R., & McClelland, J. L. (1987). Conspiracy effects of word pronunciation. *Journal of Memory and Language, 26,* 608–631.

Taylor, I., & Taylor, M. M. (1995). *Writing and literacy in Chinese, Korean and Japanese.* Amsterdam: John Benjamins.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Rose, E., Lindamood, P., Conway, T., & Garvan, C. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses in instruction. *Journal of Educational Psychology, 91,* 579–593.

Treiman, R., Tincoff, R., & Richmond-Welty, E. D. (1996). Letter names help children to connect print and speech. *Developmental Psychology, 32,* 505–514.

Treiman, R., Tincoff, R., & Richmond-Welty, E. D. (1997). Beyond zebra: Preschoolers knowledge about letters. *Applied Psycholinguistics, 18,* 391–409.

Treiman, R., Weatherston, S., & Berch, D. (1994). The role of letter names in children's learning of phoneme–grapheme relations. *Applied Psycholinguistics, 15,* 97–122.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1994). Development of reading related phonological processing abilities: New evidence of bidirectional causality from a latent variable longitudinal study. *Developmental Psychology, 30,* 73–87.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., Donahue, J., & Gasron, T. (1997). Changing

relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. *Developmental Psychology, 33,* 468–479.

Yi, K. (1998). The internal structure of Korean syllables: Rhyme or body? *Korean Journal of Experimental and Cognitive Psychology, 10,* 67–83.

Yoon, H.-K. (1997). *A study on the Hangul reading development: Acquisition of grapheme–phoneme correspondence rule.* Unpublished doctoral dissertation, Pusan National University, Pusan, Korea.

Yoon, H.-K., Bolger, D. J., Kwon, O.-S., & Perfetti, C. A. (2002). Subsyllabic units in reading: A difference between Korean and English. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy.* Amsterdam: John Benjamins.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 131,* 3–29.

Received September 8, 2008
Revision received March 31, 2009
Accepted April 21, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 948–963

0022-0663/09/$12.00 DOI: 10.1037/a0016096

A Longitudinal Analysis of Achievement Goals: From Affective Antecedents to Emotional Effects and Achievement Outcomes

Lia M. Daniels
University of Alberta

Robert H. Stupnisky
University of Manitoba

Reinhard Pekrun
University of Munich

Tara L. Haynes, Raymond P. Perry, and Nancy E. Newall
University of Manitoba

Affect and emotions are frequently seen as outcomes of mastery and performance goals, but affective experiences may also predict goal adoption. In a predictive study (N = 669 first-year college students), the authors used structural equation modeling to estimate relationships from 2 initial affective experiences to mastery and performance-approach goals, from goals to discrete emotions, and from discrete emotions to final grades in a university course while controlling for prior achievement. Representing initial affective experiences, hopefulness positively predicted mastery and performance goals, whereas helplessness negatively predicted mastery goals. Mastery goals positively predicted enjoyment, which in turn positively predicted achievement, and negatively predicted boredom, which in turn negatively predicted achievement. Anxiety was negatively predicted by mastery goals, positively predicted by performance goals, and exerted a negative predictive influence on achievement. The findings suggest that predictive relationships between goals and achievement are mediated by students' emotions. Results are discussed with regard to the importance of affect and emotions for achievement goal theory.

Keywords: achievement goals, emotion, first-year college students, achievement, structural equation modeling

Considerable evidence shows that achievement goals and affect are intricately related. However, within the literature, researchers have conceptualized affect as both an *outcome* of goal pursuit (e.g., Pekrun, Elliot, & Maier, 2006) and as an *antecedent* of goal adoption (e.g., Elliot & Thrash, 2002; Seifert, 1995). The relationships are further complicated by the multitude of definitions and operationalizations of affect and emotions (e.g., Frijda, 1993; Rosenberg, 1998). In response, several models detailing possible relationships between goals, affect, and discrete emotions have been proposed (e.g., Linnenbrink & Pintrich, 2002; Pekrun et al., 2006; Pekrun, Elliot, & Maier, in press; Seifert, 1995).

The concepts of affect and emotions used in these models pertain to general positive and negative affect, discrete emotions, and individual dispositions. *Emotions* are defined as multiple component processes composed of affective, cognitive, physiological, and behavioral elements (Scherer, 2000; e.g., for anxiety: feeling nervous, worried, increased activation, anxious facial expression). Compared with emotions, *moods* are of lower intensity and lack a specific referent (Rosenberg, 1998). Different emotions and moods are compiled in the more general constructs of positive versus negative *affect* (Tellegen, Watson, & Clark, 1999); positive affect being an omnibus variable composed of emotions such as enjoyment, pride, and satisfaction, and negative affect as an omnibus variable composed of emotions such as anxiety, frustration, and sadness (e.g., Pintrich, 2000).

These types of affective variables, and their relationships with goals, may be particularly relevant to first-year university students who find themselves in a new achievement setting that differs markedly from high school. The pressure in such novel, highly competitive learning environments is appreciable: At the extreme, approximately 27% of college freshmen do not complete their first year (Cravatta, 1997; Feldman, 2005; Geraghty, 1996). Economic and personal issues surely explain some of the attrition; however, the characteristics of a new achievement setting, such as increased pressure to excel, high demands for autonomy, and emotional instability, also contribute (Perry, 1991, 2003). Under these conditions, students' goals may be particularly susceptible to the influence of affective experiences. Likewise, students' emotions may be readily shaped by the goals they endorse in their new achievement setting. Trying to capture this sequence of events, we focused on both antecedent and outcome relationships between goals and affect or emotions in the present study (Linnenbrink & Pintrich, 2002; Pekrun et al., 2006), and their subsequent effects on first-year university students' achievement (Pekrun et al., in press).

Lia M. Daniels, Department of Educational Psychology, University of Alberta, Edmonton, Alberta, Canada; Robert H. Stupnisky, Tara L. Haynes, Raymond P. Perry, and Nancy E. Newall, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada; Reinhard Pekrun, Department of Psychology, University of Munich, Munich, Germany.

Correspondence concerning this article should be addressed to Lia M. Daniels, Department of Educational Psychology, University of Alberta, 6-102 Education North, Edmonton, Alberta, Canada, T6G 2G5. E-mail: lia.daniels@ualberta.ca

More specifically, in a structural model, first we tested whether hopefulness and helplessness, as indicators of positive and negative initial course-related affective experiences, functioned as antecedents of mastery and performance-approach goals. This argument extends from Linnenbrink and Pintrich's (2002) bidirectional model of affect and goals, which highlights the impact of affect on goal adoption. Similar to general affective states or dispositions, we expected feelings of hopefulness and helplessness to influence students' appraisals of their achievement environment, thereby affecting goal adoption (Elliot & Thrash, 2002; Linnenbrink, 2007; Linnenbrink & Pintrich, 2002). Second, following Pekrun and colleagues' (2006) model, we tested mastery and performance goals as predictors of discrete achievement emotions (i.e., enjoyment, boredom, and anxiety). Third, using their expanded goal-emotion-achievement model (Pekrun et al., in press), we extended existing perspectives on goals and emotions by considering emotions as predictors of achievement, and hence also as mediators of goal effects on achievement. Taken together, these three purposes contribute to a unique affect-goals-emotions-achievement model that informs not only the relationships among affect, goals, and discrete emotions but also their respective effects on academic achievement.

Affect as an Antecedent of Achievement Goals

In present models of affect and goals, students' affective dispositions (Elliot & Thrash, 2002) and affective classroom experiences (Linnenbrink & Pintrich, 2002) are hypothesized to influence mastery and performance goals. This relationship is underpinned by the logic that positive affective experiences facilitate the retrieval of positive self- and task-related information, whereas negative affective experiences facilitate the retrieval of negative self- and task-related information (i.e., congruency, Singer & Salovey, 1988; see also Elliot & Thrash, 2002; Linnenbrink, 2007; Linnenbrink & Pintrich, 2002). This information is then assumed to shape the extent to which students feel equipped to adopt approach-valenced goals, thus supporting the hypothesis that positive affect facilitates, and negative affect impedes, the adoption of approach goals. In general, the evidence from cross-sectional studies suggests that pleasant affect consistently relates positively to mastery goals (for reviews, see Linnenbrink & Pintrich, 2002; Pekrun et al., 2006). In contrast, the relationships between mastery goals and unpleasant affect, and between performance goals and either valence of affect, are inconsistent.

In one of only a few longitudinal investigations of the topic, Elliot and McGregor (1999) examined trait test anxiety as a possible dispositional antecedent of achievement goals. They found that anxiety positively predicted both performance-approach and performance-avoidance goals but did not predict mastery goals. Investigating general dispositional antecedents of goals, Elliot and Thrash (2002) found that positive affectivity positively predicted mastery goals and that both positive and negative affectivity positively predicted performance-approach goals. As for experimental studies testing affect as an antecedent of goals, Linnenbrink and Pintrich (2001) induced college students into positive, negative, or neutral moods. They found that students in the negative mood condition were less likely to endorse mastery-approach goals than students in the other two conditions. Mood was unrelated to performance-approach goals.

Given the paucity of studies analyzing affect as an antecedent of achievement goals, more research is clearly needed. Investigations with first-year students may be particularly relevant because university represents a new, highly competitive achievement setting in which students are likely to experience affective instability (Perry, 1991, 2003): These fluctuations could substantially impact their goal adoption. In seeking to address this issue more explicitly, we focused on two specific initial affective experiences particularly relevant to new achievement settings rather than on mood, general emotionality, or dispositional affect. Specifically, a positive initial affective experience was characterized by feeling hopeful, whereas a negative initial affective experience was characterized by feeling helpless.

According to appraisal theories of emotion, hopefulness and helplessness consist of both affective and cognitive components (Pekrun, 2006; Schunk, Pintrich, & Meece, 2008; Weiner, 1985). Feelings of hopefulness and helplessness emerge in response to appraisals of control and stability (Schunk et al., 2008). When people perceive cognitive and behavioral agency over important future goals, they feel hopeful (Snyder, 2000; Weiner, 1985). Hopefulness is a positive predictor of psychological well-being, life satisfaction, and graduation rates and a negative predictor of psychological distress and college dropout rates (Bailey, Eng, Frisch, & Snyder, 2007; Shorey, Little, Snyder, Kluck, & Robitschek, 2007; Snyder et al., 2002). In contrast, perceiving that a future goal is beyond one's control often results in feelings of helplessness (Schunk et al., 2008). Helplessness is associated with passivity, a pessimistic explanatory style, depression, and decreased agency (Abramson, Seligman, & Teasdale, 1978; Alloy, Peterson, Abramson, & Seligman, 1984; Maier & Seligman, 1976). These results imply that feelings of hopefulness and helplessness will capture students' early affective experiences on the basis of their appraisals of their new achievement setting and, thus, are appropriate goal antecedents.

In summary, different notions of affect have been considered as antecedents of mastery and performance goals; however, more research is needed on how the initial affective experiences of first-year university students influence their goals. Toward this end, we investigated hopefulness and helplessness as goal antecedents. Moreover, we tested the influence of these two initial affective experiences on goal adoption as part of a larger model including discrete emotions as goal outcomes.

Discrete Emotions as Goal Outcomes

Research investigating the relationship between mastery and performance goals and discrete emotions is becoming more pronounced and uses both cross-sectional and longitudinal data. We focused on three discrete achievement emotions, namely enjoyment, boredom, and anxiety, first, because they are frequently reported by students in achievement settings (Pekrun, Goetz, Titz, & Perry, 2002). Second, according to the control-value theory of emotion (Pekrun, 2006; Pekrun, Frenzel, Goetz, & Perry, 2007), these emotions differ systematically in terms of valence (i.e., positive vs. negative) and activation (i.e., activating vs. deactivating), both of which are assumed to be pivotal to examining the effects of emotions on achievement (Pekrun et al., 2007). Within this perspective, enjoyment is seen as a positive activating emo-

tion, boredom as a negative deactivating emotion, and anxiety as a negative activating emotion.

Third, recent empirical evidence suggests that enjoyment, boredom, and anxiety are related to mastery and performance goals. Daniels et al. (2008) used cluster analysis to identify four goal combinations (multiple, mastery, performance, and low motivation) and tested for between-cluster differences in enjoyment, boredom, and anxiety. Students in the mastery cluster (i.e., high mastery, low performance goals) reported more enjoyment and less boredom and anxiety than students in the performance cluster (i.e., low mastery, high performance). Pekrun et al. (2006, in press) investigated relationships between achievement goals and discrete emotions in German and American samples. They found that, across cultures, mastery goals positively predicted enjoyment and negatively predicted boredom even when controlling for positive and negative trait affectivity. Other research also supports a positive link between mastery goals and enjoyment (Barron & Harackiewicz, 2001; Harackiewicz, Barron, Tauer, Carter, & Elliot, 2000; Harackiewicz, Barron, Tauer, & Elliot, 2002), but boredom remains a largely neglected emotion in complex achievement settings (Fisher, 1993).

The relationships between achievement goals and anxiety are mixed. In some instances, no relationship emerged between mastery or performance-approach goals and anxiety (e.g., Pekrun et al., 2006, in press; Pintrich, 2000). We nevertheless chose to include anxiety in the present analysis for two reasons. First, as the most commonly studied achievement emotion (Zeidner, 1998, 2007), anxiety provides an informative comparison for the effects of enjoyment and boredom. Second, inconsistent findings, such as those reported next, make the relationship between approach goals and anxiety particularly important to consider.

Using the Goals Inventory (Roedel, Schraw, & Plake, 1994), which focuses on the basic dimensions of mastery and performance goals, Bandalos, Finney, and Geske (2003) found that mastery goals were negatively related to test anxiety, and performance goals were positively related to test anxiety. This pattern of results was also reported by Daniels et al. (2008), who measured mastery and performance using the Motivated Strategies for Learning Questionnaire (MSLQ) and focused on course-related anxiety. Alternatively, Sideridis (2005, Study 1) used a combination of items from preexisting scales to create a mastery scale that focused on learning and a performance scale that focused on competence and likability. He found that mastery goals were negatively related to the social alienation dimension of anxiety (see Reynolds & Richmond, 1978) but that performance-approach goals were unrelated to all dimensions of anxiety. Using the Pattern of Adaptive Learning Survey (PALS; Midgley et al., 2000), Wolters, Yu, and Pintrich (1996) and Linnenbrink (2005) found that mastery goals were unrelated to test anxiety but that performance goals positively correlated with test anxiety. This range of findings suggests that the relationship between goals and anxiety may differ depending on the operationalization of each construct. For example, when mastery and performance goals are broadly defined, as was done in Bandalos et al. (2003) and Daniels et al. (2008), it seems that mastery goals negatively predict anxiety and performance goals positively predict anxiety. Given similar conceptualizations in the present study, we expected this pattern to emerge in our results.

In summary, several studies have demonstrated relationships between mastery and performance goals and achievement-related discrete emotions. However, with the exception of research on trait anxiety and trait affectivity as antecedents of goal adoption (Elliot & McGregor, 1999; Pekrun et al., 2006, in press), previous research has not controlled for preceding affective experiences, an oversight that the present study seeks to address. Moreover, by adding academic achievement as the final step in our study, the potential for emotion to serve as a mediator between goals and achievement was tested.

Goals and Emotions Predicting Academic Achievement

Goals as Predictors of Achievement

The positive association between performance-approach goals and academic achievement is fairly robust and has been demonstrated in many empirical investigations, particularly with college students (e.g., Barron & Harackiewicz, 2001; Elliot & Church, 1997; Elliot & McGregor, 1999; Elliot, McGregor, & Gable, 1999; Harackiewicz et al., 2000). In contrast, for mastery goals in college students, it has become largely accepted that positive relationships with achievement are *uncommon* (e.g., Barron & Harackiewicz, 2001; Elliot & Church, 1997; Harackiewicz et al., 2000; Pekrun et al., in press). However, Linnenbrink-Garcia, Tyson, and Patall (2008) recently reviewed the relationships between approach-valenced goals and academic achievement in over 90 peer-reviewed articles of students in elementary school through university and found that positive relationships for mastery goals are not uncommon. Across grades, about 40% of studies revealed a positive relationship between mastery goals and achievement, less than 5% revealed a negative relationship, and about 55% of the time the results were nonsignificant (for examples of positive effects, see Church, Elliot, & Gable, 2001; Finney, Pieper, & Barron, 2004; Grant & Dweck, 2003; Rhee, Zusho, & Pintrich, 2005). Thus, although null relationships are most common, positive effects can hardly be classified as rare.

Emotions as Predictors of Achievement

According to meta-analyses, test anxiety negatively relates to achievement at all grade levels from elementary to graduate school (Hembree, 1988; Seipp, 1991). Moreover, longitudinal and experimental evidence corroborates that anxiety can negatively affect achievement (Zeidner, 1998, 2007). Aside from test anxiety, however, researchers are only beginning to consider the impact of other discrete emotions on achievement (Pekrun et al., 2002; Spangler, Pekrun, Kramer, & Hofmann, 2002; Turner, Husman, & Schallert, 2002). This empirical hesitation does not coincide with students' emotional reality: There seems to be virtually no emotion that students do not report experiencing to some extent at school (Pekrun et al., 2002). Of these other emotions, enjoyment has been found to relate positively to the quality of students' creative writing (Larson, 1989), self-regulation (Goetz, Hall, Frenzel, & Pekrun, 2006), and achievement (Daniels et al., 2008; Harackiewicz et al., 2000; Ruthig et al., 2008). Boredom, in contrast, has been negatively correlated with college students' achievement (Daniels et al., 2008; Pekrun et al., 2002; Perry, Hladkyj, Pekrun, & Pelletier, 2001; Ruthig et al., 2008).

Emotions as Mediators of the Relation Between Goals and Achievement

Most achievement goal researchers agree, "goals establish a perceptual-cognitive framework for how individuals construe and interpret achievement settings" (Elliot & Pekrun, 2007, p. 62; see also Dweck, 1986; Elliot, 1997; Pintrich, 2000). Within this framework, goals are posited to have a direct impact on achievement-relevant psychological processes, thereby influencing achievement. In this regard, several researchers suggest that emotions mediate the effects of achievement goals on grades, resulting in a goals–emotion–achievement linkage (Elliot & McGregor, 1999; Elliot & Pekrun, 2007; Linnenbrink, Ryan, & Pintrich, 1999; Pekrun et al., 2006, in press; Roeser, Midgley, & Urdan, 1996; Tanaka, Takehara, & Yamauchi, 2006).

In the present research, we combined the constructs reviewed above by estimating the relationships from students' initial affective experiences to their achievement goals, then goals to the three discrete achievement emotions, and finally discrete emotions to achievement while controlling for prior high school achievement. Research simultaneously testing nondispositional affect and emotion both as an antecedent and as an outcome of goals is lacking to date. Additionally, with few exceptions (Elliot & McGregor, 1999; Pekrun et al., in press), little research has validated the role of emotions as possible mediators of goal effects on achievement. The present study considers the antecedent–outcome relationships between goals and affect/emotions and also highlights the influence of goals and emotions on college students' achievement.

Conceptual Framework and Hypotheses

The theoretical model used for the present analysis consists of two overarching conceptual components. The first component addresses initial affective experience as an antecedent of goals and discrete emotions as a goal outcome (Elliot & Pekrun, 2007; Linnenbrink, 2007; Linnenbrink & Pintrich, 2002; Pekrun et al., 2006). The second component expands on this conceptualization and specifies the functionality of discrete emotions in (a) predicting academic achievement and (b) mediating the effects of goals on achievement (Pekrun et al., in press). Taken in its entirety, this sequential model suggests that initial affective experiences predict goals, goals predict discrete emotions, and discrete emotions predict achievement. By design, the model suggests that the relationships between constructs that occur early in the model (i.e., hopefulness and helplessness) with those that come later in the sequence (i.e., enjoyment, boredom, anxiety) are at least partially mediated by the interceding constructs (i.e., mastery, performance goals).

Students' feelings of hopefulness and helplessness, manifested in their new achievement setting, are hypothesized to begin the sequence because research shows that pleasant affect can influence adoption of the appetitive components of both mastery and performance goals (e.g., Elliot & Thrash, 2002; Kaplan & Maehr, 1999; Linnenbrink, 2007; Linnenbrink & Pintrich, 2002). Supporting this assumed relationship is the belief that "students who experience pleasant affect may perceive that they have the resources available to approach a certain outcome, making it more likely that they focus on approaching a goal for understanding or a goal for demonstrating their competence" (Linnenbrink, 2007, p. 111). Following this logic, we expected a positive relationship between initial hopefulness and both mastery and performance-approach goals. In addition, although less clearly articulated in existing models, we expected initial helplessness to relate negatively to mastery and performance-approach goals.

Pekrun and colleagues (2006) propose that achievement goals influence emotions by shaping their underlying control and value appraisals. Because students with mastery goals tend to focus on self-referent standards and the controllability and positive value of learning, we anticipated a positive relationship with enjoyment and a negative relationship with boredom (Daniels et al., 2008; Harackiewicz et al., 2000; Pekrun et al., 2006). Furthermore, we expected students high in performance goals to report heightened anxiety, indicative of their focus on outcomes and normatively defined achievement (Bandalos et al., 2003; Daniels et al., 2008; Linnenbrink, 2007). Finally, we assumed that the effects of hopefulness and helplessness on discrete emotions were mediated to some degree by goals.

For the second component relating emotions and achievement, we expected enjoyment to be positively related to students' college achievement but boredom and anxiety to be negatively related. These theorized relationships are consistent with cognitive theories suggesting that affect influences the way information is attended to and processed and, by extension, cognitive performance (Bless, 2000; Fredrickson, 2001; Levine & Burgess, 1997; Meinhardt & Pekrun, 2003; Pekrun et al., 2002). In general, it is suggested that positive-activating emotions, such as enjoyment, are associated with enhanced motivation and attention, flexible use of strategies, and self-regulation, thereby positively influencing academic achievement. As a negative-deactivating emotion, boredom likely reduces motivation, attention, and systematic processing of information, thus lowering academic achievement (Mikulas & Vodanovich, 1993; Pekrun et al., 2002). Anxiety is assumed to reduce intrinsic motivation and attention, thus correlating negatively with academic achievement (Zeidner, 1998, 2007). Finally, because goals are expected to predict discrete emotions, and discrete emotions to predict achievement, our model posits that the effects of goals on achievement were mediated to some degree by emotions.

Method

Participants and Procedure

Participants were undergraduate students in introductory psychology courses selected from the MAACH/STS longitudinal database. The MAACH/STS database is the result of a program of research involving 16 separate longitudinal studies (1992–2008) in which the academic development of college students (overall approximate $N \approx 14{,}000$) was examined at a Canadian doctorate-granting university. In creating the database, self-report data were collected twice within 1 academic year, and then merged with up to 8 years of institutional data such as grade point average (GPA; for a more complete description of the database, see Daniels et al., 2008; Perry, 2003; Stupnisky, Renaud, Daniels, Haynes, & Perry, 2008).

The two longitudinal studies based on the 1997 and 2003 cohorts were used because these were the only 2 years that included all of the variables of interest at the appropriate data collection points. Of the 796 first-year students who completed

both the Time 1 and Time 2 questionnaires, 127 were missing a response to at least one item needed for the analysis.[1] Inspection of the frequency tables for each item showed that high school average and final grade in introductory psychology were most often missing, likely because students did not grant the researchers access to their grades. The remaining missing data were on the self-report items and were likely because of respondent errors such as skipping an item or entering an invalid response. In the end, there did not appear to be any systematic reason for the missing data, and, as such, these 127 participants were excluded from the analyses in order to test multivariate normality and run bootstrap estimates in AMOS, both of which require full data (Arbuckle, 2006).

The final sample consisted of 669 first-year students who had complete data for all of the study variables. Although the cohorts were separated by 6 years, at the time each cohort completed the questionnaires they consisted of similar samples of first-year students experiencing a new achievement setting. Hence, it was expected, and found, that the relationships between initial affective experiences, goals, emotions, and achievement were consistent across cohorts.[2]

Participants completed the self-report measures on two separate occasions: The Time 1 questionnaire was completed early in the year (October), and the Time 2 questionnaire was completed 4 months later (February). Participants received research credit in their introductory psychology course and provided informed consent to the researchers to access their final grades from course instructors. Several months thereafter, first-year GPAs and final high school achievement were obtained from institutional records. The modal categorical age was 17–18 years (SD = 0.55), representing 78% of the sample (range = 17–22 years). The majority of the sample was female (68%). As for culture, the MAACH/STS database has approximately the following ethnic distribution: 70% Caucasian, 15% Asian, and 15% "other" (Hladkyj, 2002). For the 1997 and 2003 cohorts used in the present analyses, the vast majority of students (≈85%) were born in Canada and spoke English as their first language (90%).

Measures

Hopefulness and helplessness. Single-item measures were used for *hopefulness* and *helplessness*. At Time 1, students indicated the extent to which they felt hopeful and helpless about their performance in introductory psychology (see Table 1). "Hopeful" was selected to portray students' positive initial affective experience because feeling hopeful implies that students have an encouraging outlook regarding their academic development. We chose "helpless" as a descriptor of students' negative initial affective experience because feeling helpless suggests that students have a disheartening outlook regarding their progression through the course.

Mastery and performance goals. At Time 1, participants completed eight items from the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich, Smith, Garcia, & McKeachie, 1993). Four items related to mastery goals (e.g., "When I have the opportunity in my courses, I choose assignments that I can learn from, even if they don't guarantee a good grade") and four items related to performance goals (e.g., "If I can, I want to get better grades in this class than most of the other students"). In the original MSLQ, these variables were labeled *intrinsic* and *extrinsic*, respectively (Pintrich et al., 1993). However, as the field of achievement motivation shifted, it was recognized that many of the items on the MSLQ were closely related to the constructs that have become known as mastery and performance-approach goals (Harackiewicz & Linnenbrink, 2005; Kaplan & Maehr, 2007). On the basis of confirmatory factor analysis (see the Results section) and similarity with the present conceptualizations of mastery and performance-approach goals, one item was excluded from each scale (mastery α = .70, performance α = 73).

Discrete emotions. Three discrete emotions were assessed at Time 2: enjoyment, boredom, and anxiety, based on an early version of the Achievement Emotions Questionnaire (AEQ; Pekrun, Goetz, & Perry, 2005; Pekrun et al., 2002). Six items measured each emotion as it pertained to students' feelings about their psychology course. Sample items include "I enjoy learning new things" (α = .75); "When studying for this course, I feel bored" (α = .90); and, "Before I start studying material in this course, I feel tense and anxious" (α = .81).

Academic achievement. The MAACH/STS database contains several measures of academic achievement provided directly from course instructors, or institutional records, at the end of the academic year. For this study, final introductory psychology grades provided from instructors were used. The grades were reported as percentages and reflected students' performance on five or six multiple-choice tests over the two semesters. Given that all other variables in the model are course specific, the final introductory psychology grade was the most appropriate achievement outcome

[1] A total of 1,087 first-year students completed the Time 1 questionnaire, but 291 (27%) did not return to complete the Time 2 questionnaire. Many reasons contribute to the tendency for students to not complete both parts of a study, including dropping their introductory psychology course, leaving university, having completed their required research credits, forgetting about the second session, and the like. These common reasons contribute to about 30% attrition from Time 1 to Time 2 in each cohort of the MAACH/STS database. Chi-square tests for Time 1 measures (helplessness, hopefulness, mastery goals, and performance goals) comparing completers with noncompleters were nonsignificant.

[2] In order to define the sample, we tested for multigroup invariance of the measurement and structural models between the 1997 and 2003 cohorts. For the measurement models, factor loadings were constrained to be equal between the two cohorts. The results of chi-square difference tests (Byrne, 2001) are as follows: (a) mastery and performance goals unconstrained, $\chi^2(16, N = 669) = 44.58, p < .001$; fully-constrained, $\chi^2(20, N = 669) = 51.51, p < .001$; $\Delta\chi^2(4, N = 669) = 6.93, p > .05$. (b) discrete emotions unconstrained, $\chi^2(48, N = 669) = 131.03, p < .001$; fully constrained, $\chi^2(54, N = 669) = 138.72, p < .001$; $\Delta\chi^2(6, N = 669) = 7.69, p > .05$. Measurement invariance suggests that the content of each item is interpreted similarly across the cohorts. For the structural models, all structural paths were constrained to be equal between the two cohorts. The results are as follows: enjoyment unconstrained, $\chi^2(106, N = 669) = 221.47, p < .001$; enjoyment fully constrained, $\chi^2(120, N = 669) = 236.45, p < .001$; $\Delta\chi^2(14, N = 669) = 14.98, p > .05$; boredom unconstrained, $\chi^2(106, N = 669) = 211.88, p < .001$; boredom fully constrained, $\chi^2(120, N = 669) = 220.08, p < .001$; $\Delta\chi^2(14, N = 669) = 8.20, p > .05$; anxiety unconstrained, $\chi^2(106, N = 669) = 226.52, p < .001$; anxiety fully constrained, $\chi^2(120, N = 669) = 240.27, p < .001$; $\Delta\chi^2(14, N = 669) = 13.75, p > .05$. Structural invariance suggests that the relationships between the latent variables are similar across the cohorts. Taken together, these tests largely suggest the cohorts were similar, and, hence, we pooled the two into a single sample for all analyses.

Table 1
Means, Standard Deviations, and Zero-Order Correlations for the Study Variables

Variable	*M*	*SD*	Anchor	1	2	3	4	5	6	7	8	9	10
1. Hopeful	7.35	1.75	1 = *strongly disagree*, 10 = *strongly agree*										
2. Helpless	2.56	1.76	1 = *strongly disagree*, 10 = *strongly agree*	−.29*	—								
3. Mastery goals	13.02	3.39	1 = *not at all true of me*, 7 = *very true of me*	.18*	−.22*	—							
4. Performance goals	16.46	3.31	1 = *not at all true of me*, 7 = *very true of me*	.15*	.02	.25*	—						
5. Enjoyment	19.12	4.13	1 = *not at all true of me*, 5 = *very true of me*	.18*	−.19*	.31*	.15*	—					
6. Boredom	12.80	5.06	1 = *not at all true of me*, 5 = *very true of me*	−.14*	.23*	−.24*	−.09	−.39*	—				
7. Anxiety	14.90	4.85	1 = *not at all true of me*, 5 = *very true of me*	−.15*	.35*	−.14*	.08	−.02	.42*	—			
8. Final grade[a]	77.73	12.11	Percentages	.23*	−.23*	.12*	.18*	.22*	−.38*	−.22*	—		
9. High school[b]	79.03	7.94	Percentages	.09	−.14*	.02	.10*	.07	−.17*	−.15*	.41*	—	
10. GPA	3.00	0.78	4.5-point scale	.20*	−.21*	.08	.14*	.09	−.18*	−.18*	.65*	.56*	—

Note. GPA = grade point average.
[a] Final introductory psychology grade. [b] Graduating high school average.
* $p < .001$.

variable. For supplementary analyses, the model was extended in order to predict students' end-of-the-year GPAs. GPA is a widely accepted indicator of general academic performance across a range of courses and methods of evaluation (e.g., multiple-choice tests, essays, presentations, etc.). Although institutions vary somewhat, there is general agreement on the categories with our institution using the following: 4.5 = A+, 4.0 = A, 3.5 = B+, etc.

Finally, because Canadian students' performance is not measured by SATs or other standardized tests, researchers must rely on other measures of high school performance to control for preexisting aptitude differences. A substantial literature shows that high school grades are a strong predictor of college achievement (e.g., Hoffman, 2002; Zheng, Saunders, Shelley, & Whalen, 2002). Institutional records computes students' graduating high school average as a basis for university entrance requirements (i.e., English, mathematics, chemistry, and physics), thereby providing an objective measure of students' high school performance.

Gender. At Time 1, students provided demographic information, including gender (female = 454; male = 213; 2 did not indicate). This ratio is roughly equivalent to the 2:1 female-to-male ratio characteristic of cohorts from introductory psychology in the MAACH/STS database (Hall, Perry, Ruthig, Hladkyj, & Chipperfield, 2006; Haynes, Ruthig, Perry, Stupnisky, & Hall, 2006; Ruthig, Perry, Hall, & Hladkyj, 2004) and slightly exceeds the general university population (Office of Institutional Analysis, 2008).

Results and Discussion

Rationale for Analyses

We conducted our analyses in three steps. First, we created summed scales, correlated the study variables, and tested for mean-level gender differences. Second, we used confirmatory factor analyses (CFAs) to test the relationship between measured items and latent variables, a process recommended by Marsh, Byrne, and Yeung (1999) to reduce any potential measurement problems. Third, three separate structural models were estimated (one for each of enjoyment, boredom, and anxiety) and tested for gender invariance. We used separate models to minimize complexity and to obtain bootstrap confidence intervals for the indirect effect of goals on achievement through each emotion separately.[3]

We conducted measurement and structural analyses using AMOS Version 7.0 (Arbuckle, 2006) and the maximum-likelihood estimation method. We assessed quality of fit for all models by the traditional chi-square (χ^2) test, the cmparative fit index (CFI; Bentler, 1990), and the root-mean-square error of approximation (RMSEA; Browne & Cudeck, 1993). As part of the structural analyses, we also statistically tested the proposed mediational effects. The significance of these indirect effects was determined by a bootstrap method using 95% confidence intervals (CIs MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; Mallinckrodt, Abraham, Wei, & Russell, 2006; Shrout & Bolger, 2002). This method is recommended over the traditional Sobel test, which has been found to be overly conservative and to lack power even with large samples (MacKinnon et al., 2002). AMOS calculates CIs for total indirect effects using the unstandardized mean beta

[3] For interested readers, we estimated a single model including all three emotions. This model precluded testing the statistical significance of each individual emotion as a mediator of the effects of goals on achievement. Although the goodness-of-fit dropped, $\chi^2(130, N = 669) = 547.07$, $p < .001$, CFI = .90, RMSEA = .07, the paths in the single model were highly similar to those in the three separate models. Mastery goals predicted enjoyment, boredom, and anxiety (βs = .44, −.36, and −.19 respectively, ps < .01), each of which predicted achievement (βs = .12, −.21, and −.12, ps < .01), suggesting positive mediation through emotions. In contrast, performance goals positively predicted anxiety (β = .19, $p < .01$), which negatively predicted achievement, hence resulting in a negative indirect effect on achievement. Performance goals had a direct positive effect on achievement (β = .11, $p < .01$).

weight and the associated standard error from 1,000 bootstrap samples (Mallinckrodt et al., 2006; Shrout & Bolger, 2002). The indirect effect is considered significant if the CI does not include zero. Because we tested direct and indirect effects simultaneously, these calculations can reveal significant indirect effects without a significant direct effect of the predictor variable (e.g., mastery goals) on the criterion variable (e.g., achievement; MacKinnon, Krull, & Lockwood, 2000; Shrout & Bolger, 2002). The presence of mediation is indicated when the indirect effect is significant and the direct effect of the predictor on the criterion variables is nonsignificant.

Step 1: Preliminary Analyses

Correlations. Table 1 presents zero-order Pearson product–moment correlations between all study variables, several of which are highlighted here. First, hopeful and helpless were moderately negatively correlated, and mastery and performance goals were moderately positively correlated. The size of these correlations provides evidence that both the affect and the goal constructs are separate, albeit related variables, rather than opposite ends of single continuous dimensions. Second, as expected, mastery goals correlated positively with enjoyment and negatively with boredom; however, performance goals did not correlate with anxiety as was expected. Third, mastery and performance goals, as well as enjoyment, boredom, and anxiety, each correlated significantly with final introductory psychology grades. Finally, not surprisingly, high school average had a strong positive relationship with final introductory psychology grade, providing empirical support for controlling for prior achievement levels. These patterns are relatively consistent with our hypotheses based on Linnenbrink and Pintrich (2002) and Pekrun et al. (2006, in press).

Gender differences. Mean-level gender differences emerged in two instances. Female students had significantly higher high school averages than their male peers (females $M = 80.16$, $SD = 7.41$, vs. males $M = 76.67$ $SD = 8.31$), $t(375)^4 = 5.46$, $p < .001$, representing a common trend in academic achievement (Duckworth & Seligman, 2006; Mau & Lynn, 2001). Additionally, men reported more boredom than women (males $M = 14.20$, $SD = 5.13$ vs. females $M = 12.16$, $SD = 4.90$), $t(665) = 4.93$, $p < .001$. We found no mean-level gender differences for any of the other study variables.

Step 2: Measurement Models

Mastery and performance goals. We tested the mastery and performance goal measures together in a single confirmatory factor analysis (CFA). The original CFA included all items as indicators and inadequately fit the data, $\chi^2(38, N = 669) = 160.94$, $p < .001$, CFI = .91, RMSEA = .07. For both cohorts, one mastery goal item ("I prefer course material that really challenges me so I can learn new things") and one performance goal item ("I want to do well to please my family and friends") had modification indices exceeding 12.00 and cross-loaded with the other scale, suggesting that they be removed. After removing these items, the model fit improved, $\chi^2(8, N = 669) = 37.23$, $p < .001$, CFI = .97, RMSEA = .07, and the modified goal scales were retained for all further analyses (see the Appendix for item wording, loadings, and descriptive item statistics).

Discrete emotions. With six items per emotion scale, it was possible to use parceling for the emotion constructs. We used parceling in order to estimate fewer parameters, improve model fit, and reduce bias in the estimation of structural parameters (Bandalos, 2002; Little, Cunningham, Shahar, Widaman, 2002). This procedure is particularly common when researchers are interested in the relationships among latent variables more so than among individual items, as was our intention. The emotion scales had high internal reliability and demonstrated unidimensionality by exploratory factor analyses yielding single factors (eigenvalues enjoyment = 2.68, boredom = 4.02, anxiety = 3.10): Unidimensionality is a requirement for parceling (Bandalos, 2002). On the basis of the above reasons, we created three parcels for each emotion scale by summing together pairs of items that were similarly worded and correlated (range rs = .34–.71). The model fit the data well, $\chi^2(24, N = 669) = 74.28$, $p < .001$, CFI = .98, RMSEA = .06, and the parceled emotion scales were retained for all further analyses.

Step 3: Structural Models

Each of the three structural models for the different emotions was specified in terms of the same sequential process: from high school average to hopefulness and helplessness; from hopefulness and helplessness to goals; from goals to discrete emotions; and from discrete emotions to final introductory psychology grade. We estimated three additional relationships for conceptual and methodological reasons: (a) We estimated paths from hopeful and helpless to each discrete emotion to control for the direct effect of affective experiences on discrete emotions; (b) we included paths from mastery and performance goals to final course achievement to control for the direct effect of goals on achievement; and (c) we estimated a path from high school average to achievement to control for preexisting differences in academic ability. We did not expect significant direct effects for initial affective experience on emotions or for goals on achievement because these effects were believed to be mediated, at least partially, by the intervening variables of the sequential model. Because we tested the direct and indirect effects simultaneously, a nonsignificant direct effect, paired with a significant indirect effect, provided evidence in support of mediation. In total, we estimated 14 direct structural paths in each model. Two sets of indirect effects were of interest: affect-goals-emotion and goals-emotion-achievement. Finally, we correlated the residuals between hopeful and helpless and between mastery and performance goals in order to account for the interrelationships between these constructs, as has been done in recent research (Bandalos et al., 2003; Ntoumanis, Biddle, Haddock, 1999).

All three models demonstrated adequate fit: enjoyment model, $\chi^2(53, N = 669) = 164.09$, $p < .001$, CFI = .94, RMSEA = .06; boredom model, $\chi^2(53, N = 669) = 157.66$, $p < .001$, CFI = .96, RMSEA = .05; anxiety model, $\chi^2(53, N = 669) = 176.85$, $p < .001$, CFI = .94, RMSEA = .06. Overall, this suggests that the hypothesized model adequately describes antecedent and outcome relationships between hopefulness and helplessness, mastery and

[4] Degrees of freedom have been adjusted according to the Welch statistic to reflect unequal variances between groups.

performance goals, and discrete emotions. Moreover, the model seems appropriate for considering the predictive influence of goals and emotions on course-based achievement.

The standardized parameter estimates of all observed variables loaded adequately ($\beta > .40$) on their respective latent factors. However, the distribution of data for each structural model significantly departed from multivariate normality (enjoyment model multivariate kurtosis = 18.16, critical ratio [c.r.] = 11.89; boredom model multivariate kurtosis = 17.96, c.r. = 11.76; anxiety model multivariate kurtosis = 18.58, c.r. = 12.17). To test whether nonnormality inflated the significance of the regression paths, we obtained estimates for 1,000 bootstrap samples for each model. The only differences that emerged were at the third decimal place; hence, we concluded that the bootstrap parameter estimates were highly similar to those of the original model and that nonnormality did not significantly affect the accuracy of the paths (Byrne, 2001; Kline, 2005). As such, all reported results are based on the original sample.

Next, we tested each structural model for gender invariance using a chi-square difference test (Byrne, 2001). To do this, we constrained all structural paths to be equal between male and female participants (Byrne, 2001; Byrne & Watkins, 2003). The results showed no significant gender differences between models, implying equivalence of structural relationships between genders. This result supports a recent review of gender differences in motivation, wherein Meece, Bower Glienke, and Burg (2006) concluded that achievement goals do not vary systematically by gender. As such, all results are presented for the whole sample, including both genders.

Consistent with our model specifications, all three models produced highly similar relationships for all paths occurring prior to the inclusion of the discrete emotion. Because the following series of results were consistent across all three models, we present them only once. First, high school average negatively predicted helplessness (βs = $-.14$, $p < .001$) and positively predicted hopefulness (βs = .10, $p < .01$). Second, as expected, hopefulness positively predicted students' mastery and performance goals (βs = .16 and .14, ps $< .001$, respectively), confirming an association between feeling hopeful about a course and the adoption of approach-valenced goals. Third, helplessness negatively predicted mastery goals (βs = $-.23$, $p < .001$). Following are the results for the paths associated with each specific discrete emotion.

Enjoyment model. Our hypothesized direct and mediated relationships for the enjoyment model were largely supported (see Figure 1). As expected, mastery goals positively predicted enjoyment ($\beta = .38$, $p < .001$), whereas performance goals were unrelated to enjoyment. Neither hopefulness nor helplessness had a direct effect on enjoyment, suggesting that the effect of initial affect was mediated by goals. Bootstrap confidence intervals revealed that the indirect effects of hopefulness and helplessness on enjoyment were significant, suggesting that these effects were indeed mediated by goals, in line with our assumptions (see Table 2).

Enjoyment was significantly related to final introductory psychology grades ($\beta = .20$, $p < .001$), even when controlling for the influence of prior high school average on achievement ($\beta = .39$, $p < .001$). Unexpectedly, the positive predictive effect of performance goals on achievement was weak and no more than marginally significant ($\beta = .07$, $p = .08$). Finally, mastery goals did not exert a direct predictive effect on achievement either. However, there was a significant positive indirect effect of mastery goals on achievement through enjoyment. Overall, the enjoyment model explained 22% of the variance in achievement.

Boredom model. We found support for the anticipated direct and indirect relationships for the boredom model (see Figure 2). As expected, mastery goals negatively predicted boredom ($\beta = -.25$, $p < .001$), suggesting that mastery goals can protect students from this detrimental emotion. Performance goals were unrelated to boredom. Feeling hopeful had no significant direct effect on boredom, suggesting that the influence of this affect was mediated by goals. Feeling helpless, however, positively predicted boredom ($\beta = .20$, $p < .001$), suggesting that this relationship was only partially mediated by the inclusion of goals in the model. Bootstrap confidence intervals indicated that the indirect effects of both

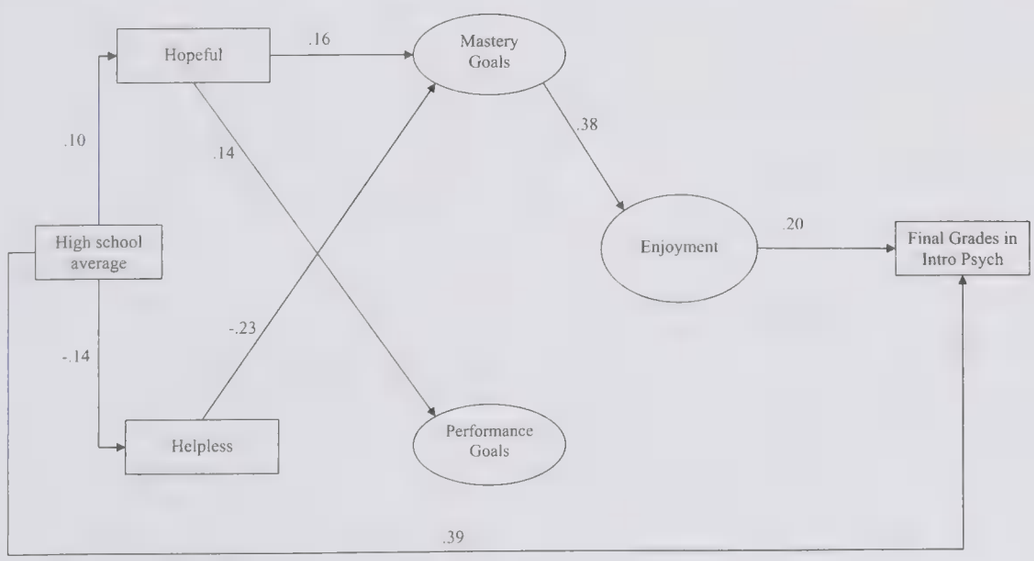


Figure 1. Structural model for enjoyment. Nonsignificant paths are not shown. All paths: $\beta = .10$, $p = .01$; βs = .11–.14, $p < .01$; $\beta \geq .15$, $p < .001$. Intro Psych = introductory psychology.

Table 2
Tests of Significance of Mediation

Independent variable	Mediating variable(s)	Dependent variable	Original sample: β standardized indirect effect	Bootstrap			
				Mean indirect effect[b]	*SE* of mean[b]	95% CI with bias correction (lower, upper)	Bias-corrected *p*
Pos. affect	Goals[a]	Enjoyment	.066	.036	.012	.017,.067	.001
Neg. affect	Goals	Enjoyment	−.086	−.047	.013	−.077,−.025	.001
Pos. affect	Goals	Boredom	−.037	−.037	.015	−.077,−.013	.004
Neg. affect	Goals	Boredom	.057	.058	.017	.028,.094	.001
Pos. affect	Goals	Anxiety	.001	.001	.012	−.023,.025	.949
Neg. affect	Goals	Anxiety	.037	.030	.013	.009,.063	.010
Mastery goals	Enjoyment	Percent	.076	.814	.251	.339,1.410	.001
Mastery goals	Boredom	Percent	.072	.754	.191	.440,1.188	.001
Mastery goals	Anxiety	Percent	.028	.288	.125	.089,.610	.004
Perform. goals	Anxiety	Percent	−.036	−.363	.123	−.671,−.180	.001

Note. Pos. = Positive; Neg. = Negative; Perform. = Performance.
[a] The AMOS bootstrap confidence intervals refer to the significance of the total indirect effect of initial affect on each emotion as mediated by both mastery and performance goals. [b] These values are based on unstandardized mean path coefficients.

feeling hopeful and helpless on boredom were significant, thus showing that these effects were mediated by goals (see Table 2).

Boredom exerted a clear negative effect on final introductory psychology grades ($\beta = -.29$, $p < .001$), even though the influence of prior high school average on course achievement was controlled for ($\beta = .35$, $p < .001$). Performance goals positively predicted final introductory psychology grade ($\beta = .09$, $p = < .05$); however, the effect was small relative to the predictive power of boredom. The lack of a direct relationship between mastery goals and achievement provides evidence of boredom mediating the predictive effect of mastery goals on achievement, an indirect effect confirmed by bootstrap confidence intervals. Overall, the boredom model explained 24% of the variance in achievement.

Anxiety model. As expected, performance goals positively predicted anxiety ($\beta = .18$, $p < .001$; see Figure 3). In addition, mastery goals negatively predicted anxiety ($\beta = -.14$, $p < .01$), suggesting that these goals can shield students from experiencing anxiety. Hopefulness did not have any significant direct or indirect effects on anxiety. In contrast, helplessness had a positive direct effect on anxiety ($\beta = .34$, $p < .001$). Bootstrap confidence intervals also indicated that the positive indirect effect of helplessness on anxiety was significant (see Table 2), suggesting partial mediation by goals.

Even when controlling for the influence of prior high school average on final course grades ($\beta = .37$, $p < .001$), anxiety significantly predicted achievement ($\beta = -.20$, $p < .001$). Performance goals had a positive direct effect on achievement ($\beta = .12$, $p < .01$) and a negative indirect effect mediated by anxiety, suggesting that anxiety triggered by performance goals may reduce the positive predictive effects of performance goals on achievement. Mastery goals did not exert a direct predictive influence on achievement. However, the indirect effect of mastery goals mediated by anxiety was significant. Overall, the anxiety model explained 21% of the variance in achievement.

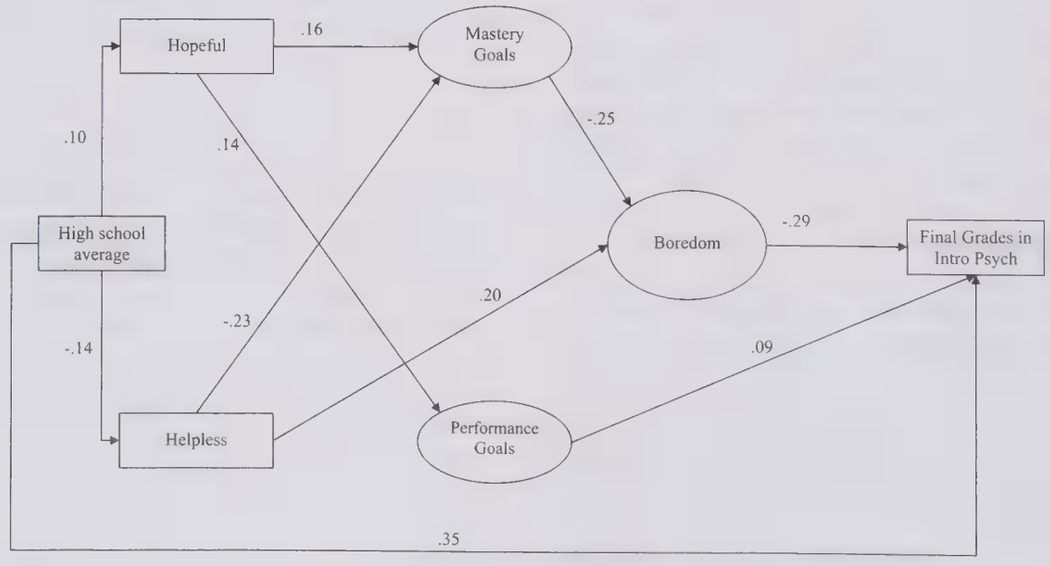


Figure 2. Structural model for boredom. Nonsignificant paths are not shown. All paths: $\beta = .10$, $p = .01$; βs = .11–.14, $p < .01$; $\beta \geq .15$, $p < .001$. Intro Psych = introductory psychology.

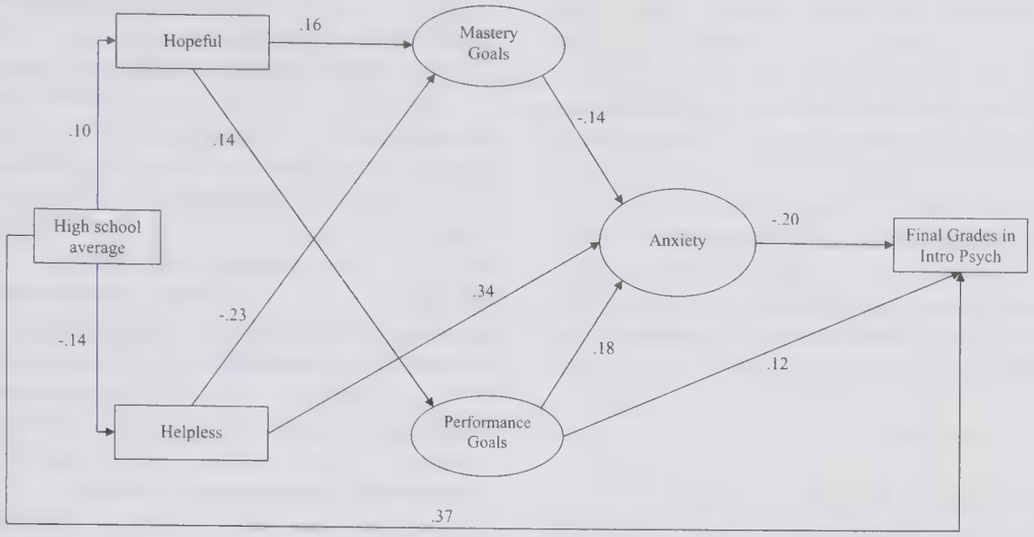


Figure 3. Structural model for anxiety. Nonsignificant paths are not shown. All paths: $\beta = .10$, $p = .01$; βs = .11–.14, $p < .01$; $\beta \geq .15$, $p < .001$. Intro Psych = introductory psychology.

Supplemental Analyses

We tested the generalizability of the model beyond introductory psychology by estimating the same models using GPAs available from institutional records. Because the only change to the model was the achievement outcome, we only anticipated changes in the paths from achievement goals and emotions to GPA. Because we measured goals and emotions at a specific level (i.e., with regard to students' introductory psychology courses), we expected their relationships with the more general GPA measure of achievement to be weaker than their relationships with final introductory psychology grade.

The models again fit the data well: enjoyment model, $\chi^2(53, N = 669) = 143.17$, $p < .001$, CFI = .95, RMSEA = .05; boredom model, $\chi^2(53, N = 669) = 143.00$, $p < .001$, CFI = .97, RMSEA = .05; anxiety model, $\chi^2(53, N = 669) = 163.02$, $p < .001$, CFI = .95, RMSEA = .06. Across all three models, no differences emerged between the original sample and the 1,000 bootstrap samples, suggesting that the paths were sufficiently accurate despite nonnormality (Byrne, 2001; Kline, 2005). Consequently, we report our results based on the original sample. The models also demonstrated gender invariance. Results that are consistent with the main models described above are collapsed across all three GPA models and presented first, followed by a more thorough presentation of changes related to the new operationalization of achievement.

As expected, based on the specification of the model, all paths *unrelated* to the achievement measure were consistent with the findings for the main models. High school average was positively related to a positive initial affective experience and negatively related to a negative initial affective experience. In turn, a positive initial affective experience positively predicted mastery and performance goals. A negative initial affective experience was negatively related to mastery goals and unrelated to performance goals. In addition, a negative initial affective experience positively predicted both boredom and anxiety. Mastery goals positively predicted enjoyment and negatively predicted boredom and anxiety, whereas performance goals positively predicted anxiety. Finally, high school average was the strongest predictor of GPA across all three models (βs = .55, .54, and .54, ps $< .001$, for enjoyment, boredom, and anxiety, respectively).

The positive relationship that emerged between enjoyment and achievement in terms of final introductory psychology grades did not persist when achievement was defined in terms of GPA. However, both boredom and anxiety continued to negatively predict achievement in terms of GPA, although the magnitude of influence was reduced (GPA: $\beta = -.14$ and $-.14$, ps $< .01$, vs. grades: $\beta = -.29$ and $-.20$, ps $< .001$, for boredom and anxiety, respectively). Mastery goals did not predict GPA, whereas performance goals positively predicted GPA in the boredom and anxiety models.

As with the main analyses, significance of mediation was determined by examining 95% confidence intervals associated with 1,000 bootstrap samples (Mallinckrodt et al., 2006; Shrout & Bolger, 2002). Accordingly, the predictive effects of mastery goals on achievement were significantly mediated by boredom (CI: lower = .95, upper = 4.40, $p < .01$) and anxiety (CI: lower = .42, upper = 2.78, $p < .001$), and the predictive effects of performance goals were significantly mediated by anxiety (CI: lower = −3.47, upper = −.71, $p < .001$).

General Discussion

The purpose of this study was to test a model in which it was hypothesized that first-year students' initial affective experiences would predict goal adoption, goal adoption would predict subsequent discrete emotions, and discrete emotions would predict achievement at the end of the academic year. Several consistencies across all three models were apparent and support the theoretical framework underpinning our model specifications (Linnenbrink & Pintrich, 2000; Pekrun et al., 2006, in press). Three of these findings are particularly important to the goals and emotion literatures. First, our study suggests that initial hopefulness positively predicted both mastery and performance-approach goals, whereas initial helplessness negatively predicted mastery goal adoption. Second, as expected, mastery and performance goals were differ-

entially related to discrete achievement emotions. Mastery goals were positively related to enjoyment and negatively related to boredom and anxiety, whereas performance goals were positively related to anxiety. Third, the model supports the premise that discrete emotions predict achievement and significantly mediate the effects of goals on achievement, both at the individual course level and more generally (i.e., course grades and GPA). Additionally, it is noteworthy that the model was equivalent for male and female students. Our discussion focuses on the extent to which these results converge with existing research and our model assumptions regarding the relationships between goals and affect/emotions, and the influence of goals and emotions on achievement.

Affect as a Predictor of Achievement Goals

The present study contributes to the literature by addressing specific initial affective experiences, rather than general traits (Elliot & McGregor, 1999; Elliot & Thrash, 2002), as goal antecedents. It appears that feeling hopeful is a positive predictor of both mastery and performance-approach goals, and thus is similar to positive dispositional antecedents in this respect. Linnenbrink and Pintrich (2002) foresaw this possibility in their model and suggested that students who feel positively may be more likely to adopt approach than avoidance goals, but may not distinguish between mastery and performance goals. Elliot and Thrash (2002) expressed a similar sentiment in their idea of "valence symmetry," which suggests that an approach temperament, consisting of positive emotionality, extraversion, and behavioral activation, will predict both mastery and performance-approach goals (see also Elliot & Pekrun, 2007).

Additionally, our results reveal that feeling helpless is negatively related to mastery goals. Linnenbrink and Pintrich (2001) similarly found that students induced into a negative mood were less likely to endorse mastery goals than those in either a positive or a neutral mood. Conflicting with this result, however, negative dispositional antecedents, such as avoidance temperament, appear to be unrelated to mastery goals (Elliot & McGregor, 1999). As such, it seems that a divergence exists between effects of helplessness as specific negative affective experiences versus general negative dispositions as antecedents of achievement goals, such that only the former are detrimental for the adoption of mastery goals. It remains an empirical question for future research, however, whether these results will generalize to affective experiences other than hopefulness and helplessness (e.g., relaxation, happiness, anger, sadness, etc.).

Achievement Goals as Predictors of Discrete Emotions

Mastery and performance goals were differentially related to enjoyment, boredom, and anxiety. As expected, and in line with the assertions of the control-value theory of emotions (Pekrun et al., 2006), mastery goals positively predicted enjoyment and negatively predicted boredom, whereas performance goals were unrelated to these two emotions. Although enjoyment and boredom differ in terms of activation and valence, they share another categorical dimension addressed by the control-value theory: *object focus*. Enjoyment and boredom are considered activity-focused emotions, meaning that they are experienced during tasks such as studying or completing an assignment (Pekrun et al., 2006, in press). Our results support the notion that because mastery goals focus students' attention on the learning process, they relate to activity-focused emotions, whereas performance goals do not.

In contrast to enjoyment and boredom, anxiety is an *outcome-focused* emotion, meaning that it is experienced in relation to the outcome of a test or an assignment. Although recent research suggests that neither mastery nor performance-approach goals relate to anxiety (e.g., Pekrun et al., 2006, in press; Pintrich, 2000), empirical findings have generally been mixed. In the present study, mastery goals were found to protect students from anxiety, whereas performance-approach goals increased students' anxiety. In explaining these relationships, the original assertions of Dweck and Leggett (1988) may be consulted. From the outset, Dweck and Leggett argued that students with mastery goals tend to interpret feedback as related to effort, which, because it can be controlled, is likely to decrease feelings of anxiety. In contrast, students with performance goals are more likely to interpret feedback as an assessment of their ability. Such threats to ability would, in turn, trigger anxiety (see also Bandalos et al., 2003, for similar results).

Achievement Goals and Discrete Emotions as Predictors of Achievement

Two sets of predictive relationships with achievement were hypothesized: from performance goals to achievement, and from each discrete emotion to achievement. Performance goals significantly predicted course-based achievement in the boredom and anxiety models as well as GPA in the supplemental analyses. In the enjoyment model, there was a marginally significant direct effect of performance goals on achievement ($\beta = .07$, $p = .08$). It is possible that the effect of performance goals on achievement did not reach significance in this instance because of an overdetermination of the dependent variable by the other variables in the model. Indeed, when the direct path from high school average to final introductory psychology grades was removed, the effect of performance goals on achievement became significant ($\beta = .10$, $p = .02$). Overall, these results are in line with the consistent finding that performance-approach goals have a positive predictive effect on course-based achievement (Barron & Harackiewicz, 2001; Elliot & Church, 1997; Elliot & McGregor, 1999; Elliot et al., 1999; Harackiewicz et al., 2000).

In addition, we found substantial relationships between the three discrete emotions and achievement. These relationships are particularly impressive given that prior achievement was controlled for in our model. Because enjoyment is a positive activating emotion, the finding that it positively predicted achievement is appropriate. The same is true of the finding that boredom, a negative deactivating emotion, negatively predicted achievement. As mentioned above, enjoyment and boredom are classified as activity-focused emotions, suggesting that the emotional consequences of enjoyable and boring activities may extend beyond the experience of the activity itself. Additionally, given that anxiety has received so much more empirical attention than boredom, it is interesting to note that the effect of boredom on achievement slightly exceeded the effect of anxiety on achievement. Moreover, both boredom and anxiety also negatively predicted generalized achievement as measured by GPA. These findings suggest that, in the future, researchers may want to examine the deleterious influ-

ence of boredom on achievement with the same rigor as the effects of anxiety.

In addition to the effects of emotions on achievement, the hypothesis that emotions would mediate the effects of goals on achievement was supported. Mastery goals predicted enjoyment, boredom, and anxiety, each of which significantly predicted achievement. The indirect effects of mastery goals on achievement were significant for each emotion as a mediator. This set of relationships suggests that mastery goals have an overall positive effect on course-based achievement through increased enjoyment, and decreased boredom and anxiety. For performance goals, there was a positive effect on anxiety, which in turn negatively predicted achievement. The indirect effect of performance goals on achievement mediated by anxiety was significant, suggesting that the effects of performance goals on achievement may be decreased by anxiety. As such, the relationships between goals and subsequent emotions may be particularly important in assessing the overall effects of goals on achievement.

Limitations, Directions for Future Research, and Implications for Educational Practice

Three limitations should be considered when interpreting the present results and in designing future research. First, due to the nature of the database, the study addressed only the approach dimensions of mastery and performance goals as assessed by the MSLQ (Pintrich et al., 1993). We fully recognize that the inclusion of the avoidance dimensions of achievement goals is important. We suggest that future research should incorporate the avoidance dimensions of mastery and performance goals and consider a multiple-goals approach to the relationships of these goals with affect, emotions, and achievement.

Second, hopefulness and helplessness were measured by single items, a process that remains common despite its associated psychometric issues (e.g., Ainley & Patrick, 2006; McGregor & Elliot, 2002; McMillan, 2008; Messick, 1995). Empirically, the correlational analysis showed substantial convergent validity between hopefulness, helplessness, the emotions, and achievement. Conceptually, feelings of hopefulness and helplessness have a clear experiential factor, which can be captured by single-item measures (e.g., Ainely & Patrick, 2006). In these ways, hopefulness and helplessness may be like several other concepts that have been adequately measured by single items, including student ratings of instructors (Abrami & d'Apollonia, 1991), self-esteem (Robins, Hendin, & Trzesniewski, 2001), course interest (Ainley & Patrick, 2006), quality of life (Zimmerman et al., 2006), self-reported health (DeSalvo et al., 2006), and job satisfaction (Wanous, Reichers, & Hudy, 1997).

Third, the self-report measures targeted students' experiences in their introductory psychology course at one university. This kind of symmetry provides a type of precision in defining variables that may be particularly essential during the initial testing of a model, but restricts the generalizability of the results (Barrett, 2005). Future research will need to test this model in other samples of students varying by courses (e.g., math, English, etc.), age (e.g., elementary, middle, high school, and college students), achievement outcomes, and culture.

Despite the limitations of the present study, the overall model specification makes a significant contribution to the investigation of goals, emotions, and achievement. Pekrun et al. (in press) proposed a goal–emotion–achievement model that we have replicated and expanded to reflect the assumptions regarding effects of initial affect on goal adoption put forward by Linnenbrink and Pintrich (2002). In total, the affect–goal–emotion–achievement model presented here is unique in its consideration of achievement goals with hopefulness and helplessness as antecedents and both emotions and achievement as outcomes. Moreover, the results of the model offer empirical support for many of the propositions forwarded by other researchers (Elliot & McGregor, 1999; Elliot & Thrash, 2002; Linnenbrink & Pintrich, 2002; Pekrun et al., 2006, in press).

In general, future research should continue to consider both the antecedents of achievement goals and their outcomes, including emotions and achievement. First, research should focus on examining other affective antecedents and other emotion outcomes, including positive-deactivating emotions such as relaxation. Second, it will be important to test for the effects of affective *states* as antecedents while controlling for the effects of *dispositional* affective antecedents. Third, the theorized mechanisms that link these constructs should be incorporated into future analyses. For example, just as Pekrun et al. (in press) suggest that emotions mediate the effects of goals on achievement, these authors also suggest that control and value appraisals mediate the effects of goals on emotions (Pekrun et al., 2006). Likewise, although emotions predicted achievement in this study, it is theorized that these effects are mediated by students' self-regulation, learning strategies, cognitive resources, behavioral engagement, and the like (Linnenbrink, 2007; Pekrun et al., 2007). Just as it will continue to be beneficial to consider the mediational role of emotions between goals and achievement, so too will researchers' understanding of the full sequence of motivational and emotional consequences of achievement goals be enhanced by considering these additional mediational mechanisms.

In addition to theory-driven advances, it is important to consider these results in light of implications for educational practice. Achievement goal theorists have focused on establishing classroom practices that support mastery goal adoption (Ames, 1992; Meece, Anderman, & Anderman, 2006). Although many of these interventions have been implemented in elementary and high school classrooms, few adjustments have been made to college classrooms. The achievement-based and performance-driven nature of the college environment may make it more difficult to implement practices that encourage the adoption of mastery goals; however, this also makes it increasingly important to do so. If college classrooms are unlikely to become more mastery oriented, then cognitive and emotional interventions may be needed to encourage students to adopt mastery goals. Haynes, Daniels, Stupnisky, Perry, and Hladkyj (2008) have shown that attributional retraining is one such intervention that promotes mastery goals relative to performance goals among college students. Moreover, the results of the present study suggest that any intervention that reduces feelings of helplessness and encourages hopefulness in students likely increases the chances of mastery-goal adoption. The first step, of course, is for instructors and students alike to become aware of the close relationships between affective experiences, achievement goals, emotions, and academic achievement, as highlighted by the findings of the present study.

References

Abrami, P. C., & D'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness—Generalizability of N = 1 research: Comment on Marsh (1991). *Journal of Educational Psychology, 83,* 411–415.

Abramson, L. Y., Seligman, M. E., & Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology, 87,* 49–74.

Ainley, M., & Patrick, L. (2006). Measuring self-regulated learning processes through tracking patterns of student interaction with achievement activities. *Educational Psychology Review, 18,* 267–286.

Alloy, L. B., Peterson, C., Abramson, L. Y., & Seligman, M. E. P. (1984). Attributional style and the generality of learned helplessness. *Journal of Personality and Social Psychology, 46,* 681–687.

Ames, C. (1992). Classroom: Goals, structures, and student motivation. *Journal of Educational Psychology, 84,* 261–271.

Arbuckle, J. L. (2006). AMOS (Version 7.00) [Computer software]. Chicago, IL: Smallwaters Corporation.

Bailey, T. C., Eng, W., Frisch, M. B., & Snyder, C. R. (2007). Hope and optimism as related to life satisfaction. *The Journal of Positive Psychology, 2,* 168–175.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9,* 78–102.

Bandalos, D. L., Finney, S. J., & Geske, J. A. (2003). A model of statistics performance based on achievement goal theory. *Journal of Educational Psychology, 95,* 604–616.

Barrett, P. (2005). What if there were no psychometrics? Constructs, complexity, and measurement. *Journal of Personality Assessment, 85,* 134–140.

Barron, K. E., & Harackiewicz, J. M. (2001). Achievement goals and optimal motivation: Testing multiple-goal models. *Journal of Personality and Social Psychology, 80,* 706–722.

Bentler, P. M. (1990). Comparative fit indexes in structural equation models. *Psychological Bulletin, 107,* 238–246.

Bless, H. (2000). The interplay of affect and cognition: The mediating role of general knowledge structures. In J. P. Forgas (Ed.), *Feeling and thinking: The role of affect in social cognition* (pp. 201–222). New York: Cambridge University Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of testing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 445–455). Newbury Park, CA: Sage.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming.* Mahwah, NJ: Erlbaum.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34,* 155–175.

Church, M. A., Elliot, A. J., & Gable, S. L. (2001). Perceptions of classroom environment, achievement goals, and achievement outcomes. *Journal of Educational Psychology, 93,* 43–54.

Cravatta, M. (1997). Hanging on to students: College student attrition rate after freshman year. *American Demographics, 19,* 41.

Daniels, L. M., Haynes, T. L., Stupnisky, R. H., Perry, R. P., Newall, N., & Pekrun, R. (2008). Individual differences in achievement goals: A longitudinal study of cognitive, emotional, and achievement outcomes. *Contemporary Educational Psychology, 33,* 584–608.

DeSalvo, K. B., Fisher, W. P., Tran, K., Bloser, N., Merrill, W., & Peabody, J. (2006). Assessing measurement properties of two single-item general health measures. *Quality of Life Research, 15,* 191–201.

Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology, 98,* 198–208.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41,* 1040–1048.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychology Review, 95,* 256–273.

Elliot, A. J. (1997). Integrating "classic" and "contemporary" approaches to achievement motivation: A hierarchical model of approach and avoidance achievement motivation. In P. Pintrich & M. Maehr (Eds.), *Advances in motivation and achievement* (Vol. 10, pp. 143–179). Greewich, CT: JAI Press.

Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 13,* 73–92.

Elliot, A. J., & McGregor, H. A. (1999). Test anxiety and the hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 76,* 628–644.

Elliot, A. J., McGregor, H. A., & Gable, S. (1999). Achievement goals, study strategies, and exam performance: A mediational analysis. *Journal of Experimental Social Psychology, 91,* 549–563.

Elliot, A. J., & Pekrun, R. (2007). Emotion in the hierarchical model of approach-avoidance achievement motivation. In P. A. Schutz & R. Pekrun (Eds.), *Emotions in education* (pp. 57–73). Amsterdam: Elsevier.

Elliot, A. J., & Thrash, T. M. (2002). Approach-avoidance motivation in personality: Approach and avoidance temperaments and goals. *Journal of Personality and Social Psychology, 82,* 804–818.

Feldman, R. S. (2005). *Improving the first year of college: Research and practice.* Mahwah, NJ: Erlbaum.

Finney, S. J., Pieper, S. L., & Barron, K. E. (2004). Examining the psychometric properties of the achievement goal questionnaire in a general academic context. *Educational and Psychological Measurement, 64,* 365–382.

Fisher, C. D. (1993). Boredom at work: A neglected concept. *Human Relations, 46,* 395–417.

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist, 56,* 218–226.

Frijda, N. H. (1993). The place of appraisal in emotion. *Cognition & Emotion, 7,* 357–387.

Geraghty, M. (1996, July 19). More students quitting college before sophomore year, data show. *The Chronicle of Higher Education,* pp. A35–A36.

Goetz, T., Hall, N. C., Frenzel, A. C., & Pekrun, R. (2006). A hierarchical conceptualization of enjoyment in students. *Learning and Instruction, 16,* 323–338.

Grant, H., & Dweck, C. S. (2003). Clarifying achievement goals and their impact. *Journal of Personality and Social Psychology, 85,* 541–553.

Hall, N. C., Perry, R. P., Ruthig, J. C., Hladkyj, S., & Chipperfield, J. G. (2006). Primary and secondary control in achievement settings: A longitudinal field study of academic motivation, emotions, and performance. *Journal of Applied Social Psychology, 36,* 1430–1470.

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M., & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology, 92,* 316–330.

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology, 94,* 562–575.

Harackiewicz, J. M., & Linnenbrink, E. A. (2005). Multiple achievement goals and multiple pathways for learning: The agenda and impact of Paul R. Pintrich. *Educational Psychologist, 40,* 75–84.

Haynes, T. L., Daniels, L. M., Stupnisky, R. H., Perry, R. P., & Hladkyj, S. (2008). The effect of attributional retraining on mastery and performance motivation among first-year college students. *Basic and Applied Social Psychology, 30,* 198–207.

Haynes, T. L., Ruthig, J. C., Perry, R. P., Stupnisky, R. H., & Hall, N. C.

(2006). Reducing the academic risks of over-optimism: The longitudinal effects of attributional retraining on cognition and achievement. *Research in Higher Education, 47,* 755–779.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58,* 47–77.

Hladkyj, S. (2002). [Effect of media and culture on moral behaviour]. Unpublished raw data.

Hoffman, J. L. (2002). The impact of student cocurricular involvement in student success: Racial and religious differences. *Journal of College Student Development, 43,* 712–739.

Kaplan, A., & Maehr, M. L. (1999). Achievement goals and student well-being. *Contemporary Educational Psychology, 24,* 330–358.

Kaplan, A., & Maehr, M. L. (2007). The contributions and prospects of goal orientation theory. *Educational Psychology Review, 19,* 141–184.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Larson, R. W. (1989). Emotions and the creative process: Anxiety, boredom, and enjoyment as predictors of creative writing. *Imagination, Cognition, and Personality, 9,* 275–292.

Levine, L. J., & Burgess, S. L. (1997). Beyond general arousal: Effects of specific emotions on memory. *Social Cognition, 15,* 157–181.

Linnenbrink, E. A. (2005). The dilemma of performance-approach goals: The use of multiple-goal contexts to promote students' motivation and learning. *Journal of Educational Psychology, 97,* 197–213.

Linnenbrink, E. A. (2007). The role of affect in student learning: A multi-dimensional approach to considering the interaction of affect, motivation, and engagement. In P. A. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 107–124). San Diego, CA: Academic Press.

Linnenbrink, E. A., & Pintrich, P. R. (2001, April). *The relation between achievement goals and affect: Moods, emotions, and directionality.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Linnenbrink, E. A., & Pintrich, P. R. (2002). Achievement goal theory and affect: An asymmetrical bidirectional model. *Educational Psychologist, 37,* 69–78.

Linnenbrink, E. A., Ryan, A. M., & Pintrich, P. R. (1999). The role of goals and affect in working memory function. *Learning and Individual Differences, 11,* 213–230.

Linnenbrink-Garcia, L., Tyson, D. F., & Patall, E. A. (2008). When are achievement goal orientations beneficial for academic achievement? A closer look at main effects and moderating factors. *International Review of Social Psychology, 21,* 19–70.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9,* 151–173.

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding, and suppression effect. *Prevention Science, 1,* 173–181.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7,* 83–104.

Maier, S. F., & Seligman, M. E. P. (1976). Learned helplessness: Theory and evidence. *Journal of Experimental Psychology: General, 105,* 3–46.

Mallinckrodt, B., Abraham, W. T., Wei, M., & Russell, D. W. (2006). Advances in testing the statistical significance of mediation effects. *Journal of Counseling Psychology, 53,* 372–378.

Marsh, H. W., Byrne, B. M., & Yeung, A. S. (1999). Causal ordering of academic self-concept and achievement: Reanalysis of a pioneering study and revised recommendations. *Educational Psychologist, 34,* 155–167.

Mau, W., & Lynn, R. (2001). Gender differences on the scholastic aptitude test, the American College test and college grades. *Educational Psychology, 21,* 133–136.

McGregor, H. A., & Elliot, A. J. (2002). Achievement goals as predictors of achievement-relevant processes prior to task engagement. *Journal of Educational Psychology, 94,* 381–395.

McMillan, J. H. (2008). *Educational research: Fundamentals for the consumer* (5th ed.). Boston, MA: Pearson.

Meece, J. L., Anderman, E. M., & Anderman, L. H. (2006). Classroom goal structure, student motivation, and academic achievement. *Annual Review of Psychology, 57,* 487–503.

Meece, J. L., Bower Glienke, B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology, 44,* 351–373.

Meinhardt, J., & Pekrun, P. (2003). Attentional resource allocation to emotional events: An ERP study. *Cognition & Emotion, 17,* 477–500.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., et al. (2000). *Manual for the Patterns of Adaptive Learning Scales (PALS).* Ann Arbor: University of Michigan.

Mikulas, W. L., & Vodanovich, S. J. (1993). The essence of boredom. *The Psychological Record, 43,* 3–12.

Ntoumanis, N., Biddle, S. J. H., & Haddock, G. (1999). The mediating role of coping strategies on the relationship between achievement motivation and affect in sport. *Anxiety, Stress and Coping: An International Journal, 12,* 299–327.

Office of Institutional Analysis. (2008). *ISBook: 2007–2008 Institutional Statistics.* Retrieved April, 22, 2009, from www.umanitoba.ca/admin/oia/media/isbook_2007_2008.pdf

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18,* 315–341.

Pekrun, R., Elliot, A. J., & Maier, M. A. (2006). Achievement goals and discrete emotions: A theoretical model and prospective test. *Journal of Educational Psychology, 98,* 583–597.

Pekrun, R., Elliot, A. J., & Maier, M. A. (2009). Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance. *Journal of Educational Psychology, 101,* 115–135.

Pekrun, R., Frenzel, A., Goetz, T., & Perry, R. P. (2007). The control-value theory of achievement emotions: An integrative approach to emotions in education. In P. A. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 13–36). San Diego, CA: Academic Press.

Pekrun, R., Goetz, T., & Perry, R. P. (2005). *Achievement Emotions Questionnaire (AEQ). User's manual.* Munich, Germany: University of Munich, Department of Psychology.

Pekrun, R. H., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37,* 91–105.

Perry, R. P. (1991). Perceived control in college students: Implications for instruction in higher education. In J. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 7, pp. 1–56). New York: Agathon.

Perry, R. P. (2003). Perceived (academic) control and causal thinking in achievement settings. *Canadian Psychology, 44,* 312–331.

Perry, R. P., Hladkyj, S., Pekrun, R., & Pelletier, S. (2001). Academic control and action control in the achievement of college students: A longitudinal field study. *Journal of Educational Psychology, 93,* 776–789.

Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology, 92,* 544–555.

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement, 53,* 801–813.

Reynolds, C. R., & Richmond, B. O. (1978). What I think and feel: A revised measure of children's manifest anxiety. *Journal of Abnormal Child Psychology, 6,* 271–280.

Rhee, C. K., Zusho, A., & Pintrich, P. R. (April, 2005). *Multiple-goals, multiple hypotheses: Reexamining the 2 × 2 achievement goal framework in introductory chemistry and psychology classrooms.* Poster presented at the American Educational Research Association Annual Meeting, Montréal, Quebec, Canada.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 27,* 151–161.

Roedel, T. D., Schraw, G., & Plake, B. S. (1994). Validation of a measure of learning and performance orientations. *Educational and Psychological Measurement, 54,* 1013–1021.

Roeser, R. W., Midgley, C., & Urdan, T. C. (1996). Perceptions of the school psychological environment and early adolescents' psychological and behavioral functioning in school: The mediating role of goals and belonging. *Journal of Educational Psychology, 88,* 408–422.

Rosenberg, E. L. (1998). Levels of analysis and the organization of affect. *Review of General Psychology, 2,* 247–270.

Ruthig, J. C., Perry, R. P., Hall, N. C., & Hladkyj, S. (2004). Optimism and attributional retraining: Longitudinal effects on academic achievement, test anxiety, and voluntary course withdrawal in college students. *Journal of Applied Social Psychology, 34,* 709–730.

Ruthig, J. C., Perry, R. P., Hladkyj, S., Hall, N. C., Pekrun, R., & Chipperfield, J. G. (2008). Perceived control and emotions: Interactive effects on performance in achievement settings. *Social Psychology of Education, 11,* 161–180.

Scherer, K. R. (2000). Psychological models of emotion. In J. Borod (Ed.), *The neuropsychology of emotion* (pp. 137–162). Oxford, England and New York: Oxford University Press.

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). *Motivation in education* (3rd ed.). Upper Saddle River, NJ: Pearson.

Seifert, T. L. (1995). Academic goals and emotions: A test of two models. *The Journal of Psychology, 129,* 543–552.

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research, 4,* 27–41.

Shorey, H. S., Little, T. D., Snyder, C. R., Kluck, B., & Robitschek, C. (2007). Hope and personal growth initiative: A comparison of positive, future-oriented constructs. *Personality and Individual Differences, 43,* 1917–1926.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7,* 422–445.

Sideridis, G. D. (2005). Goal orientation, academic achievement, and depression: Evidence in favor of a revised goal theory framework. *Journal of Educational Psychology, 97,* 366–375.

Singer, J. A., & Salovey, P. (1988). Mood and memory: Evaluating the network theory of affect. *Clinical Psychology Review, 8,* 211–251.

Snyder, C. R. (Ed.). (2000). *Handbook of hope: Theory, measures, and applications.* San Diego, CA: Academic Press.

Snyder, C. R., Shorey, H. S., Cheavens, J., Pulvers, K. M., Adams, V. H. I. I. I., & Wiklund, C. (2002). Hope and academic success in college. *Journal of Educational Psychology, 94,* 820–826.

Spangler, G., Pekrun, R., Kramer, K., & Hofmann, H. (2002). Students' emotions, physiological reactions, and coping in academic exams. *Anxiety, Stress, and Coping, 15,* 413–432.

Stupnisky, R. H., Renaud, R. D., Daniels, L. M., Haynes, T. L., & Perry, R. P. (2008). The interrelation of first year college students' critical thinking disposition, perceived academic control, and academic achievement. *Research in Higher Education, 49,* 513–530.

Tanaka, A., Takehara, T., & Yamauchi, H. (2006). Achievement goals in a presentation task: Performance expectancy, achievement goals, state anxiety, and task performance. *Learning and Individual Differences, 16,* 93–99.

Tellegen, A., Watson, D., & Clark, L. A. (1999). On the dimensional and hierarchical structure of affect. *Psychological Science, 10,* 297–303.

Turner, J. E., Husman, J., & Schallert, D. L. (2002). The importance of students' goals in their emotional experience of academic failure: Investigating the precursors and consequences of shame. *Educational Psychologist, 37,* 79–89.

Wanous, J. P., Reichers, A. E., & Hudy, J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology, 82,* 247–252.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92,* 548–573.

Wolters, C. A., Yu, S. L., & Pintrich, P. R. (1996). The relation between goal orientation and students' motivational beliefs and self-regulated learning. *Learning and Individual Differences, 8,* 211–238.

Zeidner, M. (1998). *Test anxiety: The state of the art.* New York: Plenum Press.

Zeidner, M. (2007). Test anxiety in educational contexts: Concepts, findings, and future directions. In P. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 165–184). San Diego, CA: Academic Press.

Zheng, J. L., Saunders, K. P., Shelley, I. I. M. C., & Whalen, D. F. (2002). Predictors of academic success for freshmen residence hall students. *Journal of College Student Development, 43,* 267–283.

Zimmerman, M., Ruggero, C. J., Chelminski, I., Young, D., Posternak, M. A., Friedman, M., et al. (2006). Developing brief scales for use in clinical practice: The reliability and validity of single-item self-report measures of depression symptom severity, psychosocial impairment due to depression, and quality of life. *Journal of Clinical Psychiatry, 67,* 1536–1541.

Appendix

Standardized Parameter Estimates, Wording, and Descriptive Statistics for Mastery and Performance Goals and Discrete Emotions Based on the Measurement Models

Item label	Parameter estimates: Mast[a]	Perf[b]	Error var.	Item wording	*M*	*SD*
AM1	.77		.88	I prefer course material that really challenges me so I can learn new things.	4.32	1.47
AM3	.66		1.14	In a class like psychology, I prefer course material that arouses my curiosity, even if it is difficult to learn.	5.18	1.42
AM5	—	—	—	Understanding content is most satisfying now.	5.04	1.37
AM7	.57		1.33	When I have the opportunity in my courses, I choose assignments that I can learn from, even if they don't guarantee a good grade.	3.52	1.40
AM2		.78	.82	Getting good grades in my classes is the most satisfying thing for me right now.	5.34	1.45
AM4		.92	.25	The most important thing for me right now is getting good grades so that I have a high grade point average.	5.50	1.32
AM6		.43	1.44	If I can, I want to get better grades in this class than most of the other students.	5.61	1.33
AM8	—	—	—	I want to do well to please my family and friends.	5.12	1.53
Parcel label[c]	**Enjoy[d]**	**Bore[e]**	**Anx[f]**	**Error Var**		
ENJ1	.74			1.15	I enjoy learning new things.	4.15	0.79
					Some topics are so fun that I look forward to studying them.	3.34	1.15
ENJ2	.54			2.09	After studying, I am pleased that I know more than before.	3.75	0.97
					After studying for this course, I feel calm and relaxed.	2.88	1.12
ENJ3	.77			1.35	Some topics are so enjoyable that I am very motivated to continue studying them.	3.14	1.16
					Because this course is fun for me, I study the materials more extensively than is necessary.	1.88	0.97
BOR1		.86		.88	When studying for this course I feel bored.	2.86	1.11
					The things I have to do for this class are often boring.	2.68	1.11
BOR2		.86		.98	The content is so boring that I often find myself daydreaming.	2.55	1.17
					When studying, my thoughts are everywhere else, except on the course material.	2.35	1.02
BOR3		.87		1.12	The materials in this subject area are so boring that I feel quite exhausted.	1.88	0.99
					Often I am not motivated to invest effort in this boring course.	2.10	1.15
ANX1			.77	1.53	Before I start studying material in this course, I feel tense and nervous.	2.60	1.20
					I feel queasy when I think of having to study and to do all the work.	2.31	1.13
ANX2			.80	1.10	When studying for this course, I worry that I won't be able to master all the material.	3.39	1.18
					When studying for this course, my heart beats fast because I am nervous.	1.87	0.94
ANX3			.77	1.59	While I am studying, I sometimes would like to distract myself in order to reduce my anxiety.	2.20	1.15
					When I have problems with learning the material in this course, I get anxious.	2.54	1.16

Note. AM = achievement motivation; var. = variance; dashes show the absence of estimates.
[a] mastery goals. [b] performance goals. [c] parcels of two items. [d] enjoyment (ENJ). [e] boredom (BOR). [f] anxiety (ANX).

Received March 17, 2008
Revision received April 9, 2009
Accepted April 9, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 964–981

0022-0663/09/$12.00 DOI: 10.1037/a0015906

Are SSATs and GPA Enough? A Theory-Based Approach to Predicting Academic Success in Secondary School

Elena L. Grigorenko
Yale University

Linda Jarvin
Tufts University

Ray Diffley III, Julie Goodyear, and Edward J. Shanahan
Choate Rosemary Hall

Robert J. Sternberg
Tufts University

Two studies were carried out to predict academic success in the highly competitive environment of a private preparatory school, Choate Rosemary Hall. The 1st study focused on the question of whether there are indicators beyond middle school grade-point average (GPA) and standardized test scores that might enhance the validity of measures for predicting success of students attending Choate. The results indicated the importance of taking into account aspects of self-regulated learning (SRL), such as academic self-efficacy, academic motivation, academic locus of control, and measures of the WICS (Wisdom, Intelligence, Creativity Synthesized) theoretical framework. Both sets of SRL and WICS indicators demonstrated incremental validity in predicting success at Choate. The 2nd study preliminarily evaluated the value of including indicators of aspects of the SRL and the WICS theoretical framework into the Choate admission process. The results of this study examined the utility of using quantified indicators other than middle-school GPA and standardized test scores for making admission decisions.

Keywords: academic success, secondary school, self-regulated learning (SRL), WICS (Wisdom, Intelligence, Creativity Synthesized)

Standardized tests play a major role in the selection processes for private secondary as well as for tertiary education in the United States. Most of the research literature, however, focuses on tertiary education—on tests taken during the high school years to determine entry and to predict success in college. The most commonly used tests for this purpose are the SAT and the ACT. Numerous studies have demonstrated the usefulness of the SAT and ACT as predictors of college success (e.g., Bridgeman, McCamley-Jenkins, & Ervin, 2000; Noble & Sawyer, 2002). There may be means of improving the predictions provided by these admissions tests, however.

There is preliminary evidence that when these tests are augmented with more wide-ranging measures, the predictive validity of the combined battery can be significantly higher than that of the conventional admissions tests alone (e.g., Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004; Sternberg & The Rainbow Project Collaborators, 2006). Investigators also have sought to apply modern cognitive theories to augment the prediction of performance at the graduate and professional levels of education (Hedlund, Wilt, Nebel, Ashford, & Sternberg, 2006).

In contrast to this rich literature on college-level tests, there is a much smaller body of literature on the predictive validity of secondary-level standardized tests (i.e., tests taken for admission to high schools, usually private high schools), such as the Secondary School Admission Test (SSAT) and the Independent Schools Entrance Examination (ISEE). The SSAT and ISEE are similar to the SAT in that they measure verbal, quantitative, and reading skills. The SSAT has been found to be predictive in a variety of settings, although the amount of research is limited. Specifically, the SSAT has been found to predict performance in secondary school Latin (Schuerger & Dizney, 1967) and to be useful in the identification of academically talented elementary school students (Lupkowski-Shoplik & Assoline, 1993; Mills & Barnett, 1992). When administered in modified form, it also can be useful with accommodations given to students with learning disabilities (Beattie, Grise, & Algozzine, 1983).

In this article, we investigated the dynamics of secondary-school GPA when predicted by the SSAT or ISEE and a variety of other indicators. These indicators are related to those used in past studies of college GPA, which predicted over and above the SAT and high-school grades (in particular, by Sternberg & the Rainbow Project Collaborators, 2006). The main objectives of the study

Elena L. Grigorenko, Child Study Center, Department of Psychology, and Department of Epidemiology & Public Health, Yale University; Linda Jarvin, Center for the Enhancement of Learning and Teaching, Tufts University; Robert J. Sternberg, Department of Psychology, Tufts University; Ray Diffley, III, Julie Goodyear, and Edward J. Shanahan, Choate Rosemary Hall.

Julie Goodyear is an employee of Foundation for a Greater Opportunity, which is funded by Carl Icahn; her husband is an employee of Choate Rosemary Hall.

This research was supported by the Icahn Family Foundation and Choate Rosemary Hall. The authors express their gratitude to David Francis for his comments on the growth modeling and Robyn Rissman and Mei Tan for their editorial assistance.

Correspondence concerning this article should be addressed to Robert J. Sternberg, Tufts University, Ballou Hall, 3rd Floor, Medford, MA 02155. E-mail: robert.sternberg@tufts.edu

were twofold. First, we wanted to investigate what differentiates students' performance at Choate Rosemary Hall, a highly selective college preparatory school (hereafter, "Choate" or "the school"). Each entering class at Choate is admitted on the basis of the students' middle-school GPAs, SSATs, and teacher recommendations. At admission, new students look quite comparable on these indicators. Nevertheless, the students show dramatic differentiation in performance during their tenure at Choate. Thus, factors other than academic abilities and competencies appear to be at play. For example, although Choate is a school and all of the students have been to school before, because it is primarily a boarding school and most students have not been to boarding schools before, a student's ability to adjust quickly to a new environment is also important. So, one objective was to identify indicators that might account for such differentiation in Choate GPA over time. Second, we wanted to transfer some of the psychological indicators differentiating students' academic performance at Choate to Choate's admission practices. The objective of this part of the work was to see whether students applying to Choate could be differentiated by indicators other than middle-school GPAs and standardized tests. In general, we wished to ascertain whether the selection processes could be enhanced by the use of these indicators. If so, these indicators might not only be instrumental in building the best "suited" classes for Choate but also give the admission office a new, informative, and empowering way to talk about admission with prospective families.

THEORETICAL FRAMEWORK

Two major theoretical contributions form the foundation for this work. A body of theoretical work on self-regulation (Boekaerts, Pintrich, & Zeidner, 2000), especially theories of self-regulated learning (SRL; Pintrich, 2000; Schunk, 2005), contributed to the research presented here. In addition, the framework of the WICS (Wisdom, Intelligence, Creativity Synthesized) theory developed by Sternberg (2003), an expansion of the theory of successful intelligence (Sternberg, 1996), structured the work. Below, we briefly describe key assumptions of these theories.

The SRL Perspective

The SRL perspective is characterized by four general assumptions (Pintrich, 2000). The first two assumptions of this perspective are substance based and the third and fourth are process based. First, students are active participants in the learning process. Thus, no learning process can occur that bypasses the individual characteristics of the learner, and no environment can "make" everyone learn. Second, students have the capacity, at least potentially, to monitor, control, and regulate aspects of their own cognition, motivation, and behavior; students influence, at least partially, the environment in which they learn. Third, students can set standards or goals to aim for in their learning and, while realizing these aims, can monitor their progress and regulate their cognition, motivation, and behavior. Fourth, learning environments (Assumption 1), students' individual characteristics (Assumption 2), and students' goals (Assumption 3) interact dynamically.

Our application of the SRL perspective is also enriched by a variety of other approaches. Specifically, these approaches included Dweck's (1999) theory that learners who believe that they can improve their intelligence perform better in challenging academic environments than do learners who believe their intelligence is fixed. In addition, we used ideas from Bandura's theory of self-efficacy, according to which it helps in accomplishing a task to believe in one's own ability (Bandura, 1996); also relevant are Luthar, Cicchetti, and Becker's theory of resilience, according to which people who demonstrate resilience in the face of failure have better chances of succeeding in the long term (Luthar, Cicchetti, & Becker, 2000); and Rotter's theory of locus of control (Rotter, 1990).

The WICS Theory

WICS builds on Sternberg's earlier models (for a review, see Sternberg, 2003) but differs in that it systematically synthesizes wisdom, intelligence, and creativity. According to WICS, wisdom, intelligence, and creativity synthesized provide a basis for turning out competent and responsible citizens. Such citizens are expected to use (a) creativity to generate new ideas and problems as well as possible solutions to the problems, (b) analytical intelligence to evaluate the quality of these solutions, (c) practical intelligence to implement decisions and persuade others of their value, and (d) wisdom to ensure that these decisions help achieve a common good over the long and short terms. Thus, we argue that a predictive assessment should involve an evaluation of not only cognitive abilities—broadly defined as analytical, practical, and creative—but also elements of responsible reasoning and moral judgment that are integral to wisdom (Sternberg, 2003). This theoretical approach framed our research with regard to what kinds of competencies are essential to student success.

EMPIRICAL FRAMEWORK

The research presented here was structured around the question posed in the title of this article: Are SSATs and middle-school GPA enough for predicting academic success in secondary school, or could and should they be augmented to improve prediction? The SRL and WICS theoretical approaches framed our research with regard to how the competencies delineated earlier needed to be measured. Specifically, we decided to measure relevant competencies through (a) students' performance on competence-related tasks; (b) students' own appraisals of their academic goals, motivation, and their awareness of the learning environment and their academic success in it; and (c) teachers' comments on how the self-regulatory processes of their students change over the period of the freshman year. The work was conducted with multiple samples over a number of years. Each of the studies had a specific objective; each contributed to answering a specific question in addition to the overarching research question. Correspondingly, we present these studies separately and then provide a general discussion of the results.

STUDY SETTING

All research was carried out at Choate (Wallingford, Connecticut). It is a secondary-level boarding school that was initially founded in 1890 as Rosemary Hall (for girls), with Choate School (for boys) following 6 years later. Rosemary Hall was created for the daughters of families who sought a more intellectually stimu-

lating academic program for their daughters than would have been available for girls at home. Choate, like other boarding schools at the time, was created to prepare the sons of well-to-do families to be civic leaders. The schools evolved a great deal through the middle of the century, and then even more so when they merged in 1972. Although transformed, the school continued to include academic excellence and civic responsibility as foundations of their program. However, it has since broadened its outreach to become a geographically, economically, and culturally diverse secondary institution. Choate Rosemary Hall is a school of 840 students, 640 boarding students from 41 states and 33 countries, and 200 day students. One of the top academic boarding schools in the country and one of the most selective, it continues to attract and educate bright and motivated students. With 240 different courses and advanced placement courses in 25 different areas, it provides its students with a rich and challenging academic environment. With 80 interscholastic teams and 60 extracurricular clubs, life outside the classroom is also active. The college admission focus begins early and is competitive. The academic, extracurricular, and social facets of Choate life are advanced. It is a privilege to attend this school for its resources are deep and broad. Choate's admission goal is to ensure that the students it accepts are hungry for these opportunities and that they will adjust to and flourish in the environment. Finding students who will thrive rather than flounder is important for everyone involved (i.e., for students, their families, teachers, and administrators). By providing a challenging, thoughtful, and varied secondary school experience, Choate hopes to prepare its students to be leaders in their future fields and communities.

Study 1

Study 1 comprises a pilot study that included a smaller group of Choate students (hereafter, Pilot Study) and a main study (hereafter, Main Study) that included a larger group of students constituting a whole class. All participation in this research was voluntary.

Pilot Study

The sample in this study comprised high-ability students from lower socioeconomic strata selected to enter the school on scholarship support provided by the Icahn Charitable Foundation and/or the Foundation for a Greater Opportunity. The Icahn Scholars Program identifies motivated and highly able middle-school students from disadvantaged backgrounds and provides 10–18[1] of them each year with a fully funded Choate education. Scholars are identified by teacher and parent recommendations, middle-school GPAs, standardized test scores (SAT, SSAT, and ISEE scores), and a semistructured interview conducted by a representative of the Foundation(s).

The importance of programs such as the Icahn Scholar Program is difficult to overstate. In a global world in which exposure to education and sources of rapidly accumulated information will only continue to be critical to participation, and in which some underresourced populations never move out of a 15-block radius, the benefits of bringing an Icahn Scholar to Choate are many. The child learns and lives in an environment that has crystallized the country's (and to some degree, the world's) highest standards of balanced academic and nonacademic education. The Icahn Scholar is introduced to, and introduces his or her family and often friends and neighbors to, this broader world. Additionally, because the Icahn Scholar Program is not entirely composed of students of color, these students refute the myth that poverty is only a factor for families of color. In turn, the graduates bring with them their Choate experiences as they move on to college and beyond. When one of the first Icahn Scholars graduated and moved on to Harvard, she wrote, "I find I am not homesick. I am Choate-sick."

On another level, the non-Icahn Scholar students at Choate learn about very different home environments, as they live with Icahn Scholars from different backgrounds (a dangerous inner city, a Navajo reservation), and this broadens the horizons of the more resourced students. As a result, when Choate graduates go into the marketplace, they are more informed, more empathetic, and more diversified citizens.

This research originated as a project to determine which qualities were important for the Icahn Scholars' academic success. When expanded to include the entire Choate student body, the results were at least equally valid. Although Choate is unusual in its cost, it is similar to any other high school that prides itself on educating students to participate in competitive educational and labor environments. Although not all high schools may have a preponderance of motivated, bright students, all high schools provide advanced courses for students. As No Child Left Behind continues to stress test scores, all schools will become more achievement driven and more focused on test results and college credits. As schools look to change their students into more academically focused populations, the results from this work can provide information on what skills to teach to increase achievement. This would be universally beneficial to all schools—to schools already educating bright, motivated students and to schools wanting to give their students the skills to become more successful students. Although this work centered on an admission process, the results can be extrapolated into classroom use to enhance learning skills and adaptability to whatever new academic environment needs to be faced throughout one's educational pathway.

Method

Participants

Over the 3 years of the Icahn Scholars Program's existence prior to this study, 55 students were accepted for the program on the basis of their school grades and standardized test scores, parent and teacher recommendations, and results of individual interviews, conducted by one of the authors. However, when admitted, some of the Icahn students did not fare as well academically as the Choate admission officers had hoped; such "unfulfilled hopes" are also true for the student population at Choate, in general. Some students do not do as well in their new school as their admission materials would predict; others, of course, do better than expected, based on their standardized scores. This research was motivated by our desire to understand the causes of this diversification of performance, which occurred among the Icahn Scholars, who,

[1] Numbers vary with available funding.

prior to their arrival at Choate, appeared to be of roughly comparable ability and motivation. Specifically, within the first semester at Choate, the students demonstrated a high level of variation in their Choate GPAs. First, some students (n = 4, out of 55 across 3 years of admissions) were not able to complete their studies and left the school. Second, whereas the pre-Choate[2] GPA for the 55 Icahn Scholars admitted prior to the beginning of this research was at a mean value of 3.84 (out of 4) with a standard deviation of 0.30, their postadmission GPAs at Choate varied dramatically. The mean value was 2.89 with a standard deviation of 0.68 in their first trimester at Choate. Thus, the question was whether a battery of assessments could be developed and administered that would be predictive of academic success for the Icahn Scholars in the Choate environment so that the quality of admission decisions could be improved.

The full sample of participants in this study included 51 students, of whom 24 (47%) were girls and 27 (53%) were boys. All were approximately 180 months (15 years) old at the year of admission. The majority came from diverse ethnic groups of color; approximately 35% were White students from low-SES backgrounds.

Procedure

Students were evaluated while in summer school at Choate, prior to the beginning of their first fall semester at Choate. All assessments were group administered and took approximately 40 min to complete. All materials were preprinted at the Psychology of Abilities, Competencies, and Expertise (PACE) Center[3] and administered and scored by the PACE research team. The Icahn Scholars were debriefed on the purposes of the work, which were explained as attempting to improve the selection procedures into the program and to maximize the likelihood of the Icahn Scholars' succeeding at Choate. After contributing their time to this research, the students were offered treats (pizza and/or dessert).

Measures

Self-reports. In developing the materials for this study, three limitations surfaced. First, the group of students was fairly homogeneous with regard to their middle-school GPAs and their standardized test scores. Thus, it was not anticipated that any analytical ability measures would greatly differentiate among them. Second, the power of the sample was fairly low (n = 51), limiting the ability to introduce a large number of measures in addition to middle-school GPA and the three SSAT indicators (Verbal, Quantitative, and Reading). Third, the school was interested in enhancing its ability specifically to predict Choate GPAs. As a result of these constraints, a decision was made to focus on self-reported characteristics within the SRL theoretical framework, assessing students' (a) self-awareness of their individual characteristics as learners and (b) capacity to regulate their learning through their academic self-esteem, self-efficacy, locus of control, and motivation. For simplicity, these measures are referred to collectively as PACE measures (PACE Battery). The devised self-report scales were as follows.

The Self-Esteem Scale captured students' perception of their own individual characteristics (e.g., cognitive abilities, personality traits, skills in dealing with peers and adults). An illustration[4] of an item typical of this scale is "You think you are usually intelligent." The scale had nine items, and its internal consistency reliability was acceptable (Cronbach's α = .85).

The Academic Self-Efficacy Scale included five items. An illustration of an item typical of this scale is "I can plan my time effectively to get my work done." The Cronbach's alpha for the scale was .75.

The Academic Locus of Control Scale included two items, whereby the students attributed their academic successes and failures either to themselves or to other forces. The Cronbach's alpha was .61.

The Intrinsic/Extrinsic Academic Motivation Scale was designed to grasp students' balance of motivation as driven by their internal forces (e.g., interest) versus external rewards (e.g., grades). An example of an item typical of this scale is "It is important to get the best grades in your class." The scale had 11 items, and its internal consistency was acceptable (Cronbach's α = .78).

For all scales, students were asked to rate their responses on a 1–7 scale ranging from 1 (*strongly agree*) to 7 (*strongly disagree*.)

School reports. The school provided (a) demographic data (gender, age, ethnicity, and citizenship); (b) grades for all subjects the students took during their 4 years at Choate, summarized as GPAs for the fall, winter, and spring trimesters for 4 consecutive years (i.e., 12 time points); and (c) preadmission data consisting of students' middle-school GPA, their standardized test results (i.e., three subscores—Verbal, Quantitative, and Reading—of the SSAT).[5]

Results

This study (Pilot Study) was aimed at understanding the contributions of middle-school GPA, SSAT/ISEE scores, and PACE-designed measures to predicting success at Choate, as captured by Choate GPA. Correspondingly, we present the results in three blocks. First, we present descriptive statistics. Second, we present growth analyses of Choate GPA. Third, we present summative regression analyses predicting average GPA while at the school.

Descriptive Statistics

Table 1 presents descriptive statistics for the GPA indices used in this study. As for other relevant indicators, the means and standard deviations (shown in parentheses) for the SSAT indices were 68.0 (24.4), 77.6 (19.9), and 78.2 (16.5) for Verbal, Reading,

[2] This term is used interchangeably with the term *middle-school GPA*.

[3] The PACE (Psychology of Abilities, Competencies, and Expertise) Center was based at Yale University and in 2006 moved to Tufts University. All new measures developed for this work were developed at PACE and, therefore, are referred to collectively as PACE measures/indicators or as the *PACE Battery*.

[4] All assessment items used in this work belong to Choate. Therefore, here we use examples illustrating, but not sampling, the actual items used in this research. For more detail, please contact Raymond Diffley, III, Director of Admission, Choate Rosemary Hall, rdiffley@choate.edu

[5] During the course of this work, the SSAT changed its scaling rules, so the raw scores differ for different studies presented here. Please see http://www.ssat.org

Table 1
Means and Standard Deviations for Indices Used With Icahn Scholars in Study 1, Pilot Study

	GPA			
	Pre-choate		Choate	
Indicator	*M*	*SD*	*M*	*SD*
Time				
Baseline	3.84	0.30		
Freshman year				
Fall			2.78	0.68
Winter			2.91	0.71
Spring			3.00	0.66
Sophomore year				
Fall			3.06	0.58
Winter			3.09	0.62
Spring			3.07	0.64
Junior year				
Fall			3.11	0.60
Winter			3.08	0.63
Spring			3.23	0.65
Senior year				
Fall			3.30	0.51
Winter			3.32	0.59
Spring			3.24	0.64

Note. GPA = grade point average.

and Quantitative, correspondingly. The descriptives for PACE SRL measures were 3.44 (1.19) for self-esteem, 3.35 (0.82) for Academic Motivation, 2.54 (0.95) for Academic Locus of Control, and 2.70 (.092) for Academic Self-Efficacy. For the outcome variable(s) of interest, the 12 indices of Choate GPA, there were no gender differences. There were asystematic ethnicity-related group differences at different time points, when ethnicity was classified into two groups (White and students of color, as captured by the ethnicity status variable). Specifically, White students slightly outperformed students of color (for 7 out of 12 data points, with p values ranging from .048 to .008). The repeated measures analyses of variance indicated the presence of the main effect of ethnicity status, $F(1, 45) = 6.55$, $p < .01$, $\eta^2 = .13$, although there were no interaction effects with the time variable. In addition, average Choate GPA differed for the two groups with regard to their ethnicity status (3.32 vs. 2.82), $F(1, 45) = 7.40$, $p < .01$, with White students outperforming students of color. On the basis of these results, we included the variable ethnicity status but not gender in subsequent analyses.

Conditional Growth in Choate GPA: What Matters?

As is apparent from the data presented in Table 1, Choate GPAs appear to increase over the period of time at school, but this growth is characterized by some fluctuation. To investigate the changes in Choate GPA over time in the group of Icahn Scholars, we fitted two-level growth models, where Level 1 was represented by growth trajectories of Choate GPA, and Level 2 was represented by student-level data, specifically, ethnicity status, pre-Choate GPA, SSAT indicators, and PACE SRL indicators. To complete these analyses, we used hierarchical linear modeling (HLM) software, Version 6 (Raudenbush, Bryk, Cheong, & Congdon, 2000).

First, an unconditional linear growth model was fitted using Icahn GPAs over the 12 trimesters at Choate. The results confirmed the dynamics in GPAs evident from Table 1: There was linear growth in Icahn GPAs over time, with the starting point at approximately 3.05 and a growth rate of .009 GPA per month or .027 GPA per trimester, $T(468) = 6.28$, $p < .001$. Now, what variables among those available in this study can predict either the starting point or the growth rate of Choate GPAs?

To answer this question, we fitted a linear growth model of GPAs at Level 1 and compared whether there were significant differences in rates of growth if the growth were conditioned on the collected indicators (Level 2). All models were fitted separately for separate indicators so that the variables specified at Level 2 were defined as predictors of both the intercept and the slope of growth. The time at Choate was captured by academic months at Choate (with summer months not counted). Table 2 presents the outcome of these analyses.

A number of observations can be drawn from Table 2. The models differed in terms of the impact of specific variables on the starting GPA value and GPA growth rate. Specifically, the intercept was influenced by (a) ethnic status, with students of color starting at a lower GPA than did White students ($p < .01$), with students of color approximately .5 GPA points behind, (b) middle-school GPA, so that students with higher pre-Choate GPA started at higher GPA at Choate ($p < .001$), with one unit of pre-Choate GPA, resulting in approximately.9966 units of Choate GPA; (c) SSAT Quantitative (SSAT-Q; $p < .05$), with one SSAT-Q unit predicting .0121 units of Choate GPA; (d) SSAT Reading (SSAT-R; $p < .05$), with one SSAT-R unit predicting .0083 units of Choate GPA; and (e) Academic Self-Efficacy ($p < .005$), with a unit of stronger efficacy predicting .2909 units of Choate GPA. Only three variables affect the rate of growth in GPA: ethnicity status ($p < .05$), with students of minority background showing slower growth; SSAT Verbal ($p < .001$), with students with higher SSAT scores showing slower growth rates; and Academic Locus of Control ($p < .001$), with students with an external locus of control exhibiting a higher rate of GPA growth.

To summarize these results in a single model, we fitted a complex equation in which Choate GPA intercept and slope were predicted by those variables whose contributions were statistically significant in single-predictor models (see Table 2). Fitting such a combined model with multiple variables resulted in selected variables losing their significance. Dropping these statistically insignificant variables demonstrates no loss of fit, (χ^2difference with 5 *df* was 5.46, *ns*). Therefore, here we discuss only the parameters from that final reduced model. According to that model, two variables influence the GPA intercept (starting value). These variables were middle-school GPA ($p < .005$), with one unit of pre-Choate GPA predicting .8022 units of Choate GPA, and Academic Self-Efficacy ($p < .01$), with one unit of self-efficacy predicting .2153 units of Choate GPA. The rate of growth was predicted by three variables, ethnicity status ($p < .005$), with minority students gaining GPA at a lower rate; SSAT Verbal ($p < .001$), with higher scores associated with slower growth in GPA; and Academic Locus of Control ($p < .05$), with more externally oriented students exhibiting higher rates of growth.

Table 2
Change in Choate Grade Point Average (GPA; Study 1, Pilot Study)

Model	Intercept (starting value)			Slope (growth rate)		
	Coefficient	*T*	*p*	Coefficient	*T*	*p*
Single-predictor model						
Parameter model						
Ethnicity status	−.4681	−2.70	.010	−.0063	−2.22	.026
Pre-Choate GPA	.9966	3.77	<.001	−.0034	−0.71	.481
SSAT Verbal	.0063	1.73	.091	−.0002	−4.06	<.001
SSAT Quantitative	.0121	2.27	.029	−.0000	−0.55	.580
SSAT Reading	.0083	2.18	.035	−.0001	−1.12	.269
Self-esteem	−.0786	−0.94	.354	.0004	0.34	.738
Academic Self-Efficacy	−.2909	36.97	.003	.0008	0.50	.619
Academic Locus of Control	−.0729	−0.74	.462	.0048	3.37	.001
Academic Motivation	−.1124	−0.97	.339	−.0004	−0.24	.813
Multiple-predictor (combined) model						
Full model						
Ethnicity status	−.0423	−0.26	.800	−.0090	−3.19	.002
Pre-Choate GPA	.6729	2.50	.017			
SSAT Verbal				−.0002	−3.79	<.001
SSAT Quantitative	.0086	1.89	.066			
SSAT Reading	.0036	0.92	.361			
Academic Self-Efficacy	−.1911	−2.23	.031			
Academic Locus of Control				.0030	2.04	.041
Reduced model						
Ethnicity status				−.0090	−3.16	.002
Pre-Choate GPA	.8022	3.13	.004			
SSAT Verbal				−.0002	−3.78	<.001
Academic Self-Efficacy	−.2153	−2.56	.015			
Academic Locus of Control				.0030	2.03	.042

Note. SSAT = Secondary School Admission Test.

Predicting Averaged Choate GPA

The growth model analyses demonstrated the importance of indicators from all types of variables considered in this work—previous GPA, standardized tests, and indicators of SRL. Some of these variables were important predictors of the starting point at Choate, whereas others predicted the dynamics of growth in GPA. In the analyses that follow, we attempted to predict the averaged Choate GPA. Such averaged GPAs are used for evaluation of college applications and, subsequently, for predicting college GPA.

The results of these analyses are shown in Table 3. Two regression equations were fit.[6] The first regression included only the SSAT and the PACE SRL indicators (see the top portion of Table 3). Collectively, SSAT indicators explained about 15% of the variance in Choate GPA, but not a single standardized coefficient was statistically significant (the model's p value was .07). The introduction of PACE SRL indicators explained approximately an additional 17% of the variance in school GPA, generating a final R^2 of .32, with $p < .05$ for the model. The second regression included middle-school GPA (see the bottom portion of Table 3). Collectively, SSAT indicators and middle-school GPA explained approximately 34% of the variance (model $p < .005$). The addition of the PACE SRL indicators added about 8% more of the explained variance, bringing the squared correlation to .42 (model $p < .008$). When the reduced regression was fit, which included only the statistically significant coefficients from the analyses presented earlier (i.e., pre-Choate GPA and Academic Self-Efficacy), the model explained 34% of the variance ($p < .001$), of which 24.6% was attributed to middle-school GPA, and 9.6% was attributed to Academic Self-Efficacy. Both coefficients were statistically significant ($\beta = .41, p < .01$ and $\beta = -.32, p < .05$, for pre-Choate GPA and Academic Self-Efficacy, respectively, with higher pre-Choate GPA and higher Academic Self-Efficacy predicting higher Choate GPA).

Discussion

The results of this study suggest that middle-school GPA and standardized test results, although informative in predicting high-school GPA, predict, at least in this sample, only a portion of the variance in Choate GPA. Additional indicators capturing characteristics of SRL are informative not only in increasing the prediction of the absolute values of GPA but also in predicting the rate of growth in GPA across the 12 trimesters (or 36 academic months) at Choate.

Although interesting, these observations are preliminary because of the small size and consequent lack of generalizability of the studied sample of Icahn scholars. In addition, the repertoire of

[6] Of note is that the inclusion of the ethnicity status variable did not improve the R^2. Because the obtained coefficient was not significant, this variable was omitted from analyses.

Table 3
Incremental Prediction of High School GPA of Icahn Scholars Using SRL-Related Measures (A) Above and Beyond SSAT and (B) Above and Beyond SSAT and Pre-Choate GPA (Study 1, Pilot Study)

Measure	Step 1	Step 2
A		
SSAT		
Verbal	.084	.043
Quantitative	.261	.267
Reading	.183	.117
PACE SRL indicators		
Self-esteem		.229
Academic Self-Efficacy		−.519**
Academic Locus of Control		−.025
Academic Motivation		.019
R^2	.159	.324
B		
SSAT		
Verbal	.024	.003
Quantitative	.182	.205
Reading	.198	.144
Pre-Choate GPA	.437**	.340*
PACE SRL indicators		
Self-esteem		.146
Academic Self-Efficacy		−.379*
Academic Locus of Control		−.009
Academic Motivation		.006
R^2	.339	.421

Note. Entries are standardized beta coefficients. GPA = grade point average; SRL = self-regulated learning; SSAT = Secondary School Admission Test; PACE = Psychology of Abilities, Competencies, and Expertise.
* $p < .05$. ** $p < .01$.

measurements was limited to self-reports only. Yet, this Pilot Study formed the foundation for our continuing the work.

Main Study

The main objective of this study was to capitalize on the preliminary findings from the Pilot Study and overcome its three main limitations by recruiting a (a) larger and (b) more representative group of participants, and by (c) including not only self-reports but also performance-based measures. In the Main Study, we worked with a whole freshman class of newly admitted Choate students.

Method

Three parameters of this study are essential in the description of its methodology. First, all of the assessments are theory driven (see the introduction). Second, to maximize the information gained from these assessments, the assessments were administered to an entire class of freshman students while they were adjusting to the environment at Choate; thus, a multiple time point longitudinal investigation of their adjustment to new learning challenges was carried out. Third, a multitrait/multimethod assessment approach was used in which different traits were assessed, different informants (students, teachers, and the school itself) were engaged, and different methodologies (self-report, rating scales, and maximum performance assessments) were used.

Participants

All freshman boys and girls (N = 152, 76 boys and 76 girls; mean age = 176.8 months, SD = 5.3) admitted to the school for the 2005–2006 academic year were asked to participate in this study. There were 95 (62.5%) White students; the rest (n = 57, 37.5%) were students of color.

In addition, 20 teachers employed by the school full time were asked to evaluate the students' adaptation to the school's environment using a structured evaluation form. Among the participating teachers, there were 8 women and 12 men (ranging from 25 to 72 years of age, mean age = 48.1) who had been teaching at the school anywhere from 2 to 50 years (mean length of teaching at the school = 14.95 years); the majority of teachers were White.

Procedure

Students were evaluated three times: in early December (n = 149), late January (n =152), and mid-April (n = 138) of their freshman year. All assessments were group administered and took approximately 90 min to complete. All materials were preprinted at the PACE Center and administered and scored by the PACE research team. After contributing their time to this research, the students were offered treats (pizza and/or dessert).

Teachers evaluated students five times, with the following due dates for the evaluation forms: early December, late January, early March, late April, and late May of their freshman year. All evaluations were completed by the teachers individually, at their own pace and in locations of their choosing.

Measures

Eight different assessments were administered to the students, and one survey was administered to the teachers; these assessments, collectively, were referred to as the PACE Battery. Some components of the battery included improved versions[7] of the self-reports from the Pilot Study; others were newly developed. As in the Pilot Study, relevant information was obtained through students' school records.

WICS constructs. Driven by the WICS theory, the following constructs were evaluated:

1. Analytical competence was assessed primarily through students' preadmission SSAT scores and preadmission grades (pre-Choate GPA).

2. Practical competence was assessed primarily through an assessment of tacit knowledge of the school's environment (the School Life Inventory) and a set of practical reasoning tasks.

[7] Specifically, we expanded the number of items for each scale, adding new statements while preserving the internal consistency of the scales.

3. Creative competence was assessed primarily through a creative writing task and by an assessment of the ability to deal with novelty (involving selective encoding and recombining of information while applying scientific knowledge to solving problems, i.e., the scientific reasoning task).

4. Ethical reasoning and wise reasoning were assessed using scenarios presenting ethical dilemmas in which students were asked to analyze an ethically charged situation, find a solution to the situation, attribute blame for a particular event, and state their reasons.

These tasks were not designed as pure measures of particular competencies. For example, the scientific reasoning task called for analytical competence when the problem needed to be identified from the description and a proper element of scientific knowledge had to be evoked. Yet, it called for creative competence when the information needed to be selectively encoded from the mass of information presented and the novelty of the problem needed to be addressed so that the problem could be reformulated and specific knowledge invoked to solve it.

SRL constructs. In addition, driven by the SRL perspective, students' self- and teacher-based comments on their academic goal orientation, motivation, and appraisal were recorded. These items were devised to capture students' self-perceptions at different stages of learning, namely, the planning of learning tasks, the monitoring and controlling of performance, and reflection on their accomplishments.

Self-reports. Self-reports were extended and improved versions of the scales were used in the Pilot Study (see above).

The Academic Self-Efficacy Scale included 12 items, on each of which students rated themselves on a scale ranging from 1 (*strongly disagree*) to 9 (*strongly agree*). Items contained statements regarding students' views of their own capabilities to produce desired levels of performance (e.g., "I can cope with any homework assignment"). Cronbach's αs were .77, .81, and .80 for Times 1, 2, and 3, respectively.

The Academic Locus of Control Scale included 16 items, on each of which students rated themselves, or rather their views of the source of control over their academic success at Choate, on a scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). An example of an item typical of this scale is "I can control my grades at school." Cronbach's αs were .74, .76, and .71 for Times 1, 2, and 3, respectively.

The Intrinsic/Extrinsic Academic Motivation Scale included 12 items, on each of which students rated themselves on a scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The items were designed to capture the students' perception of the origin of academic motivation, which can be placed internally (i.e., originating within the student, intrinsically) or externally (i.e., originating from the students' context, extrinsically). Cronbach's αs were .79, .83, and .82 for Times 1, 2, and 3, respectively.

Performance tasks. All tasks were new and developed for the purposes of this research.

The School Life Inventory included 35 items administered in two versions that capture different aspects of the implicit rules of Choate life. The items presented a general scenario characteristic of various aspects of life at Choate (e.g., working on one's homework assignments while there is a birthday party in the dorm) and various choices a student can make in such situations (e.g., stop working on the homework right away and join the party, finish the homework and then join the party, go to the party for a few minutes and then leave to finish the homework, skip the party, and so forth). Each item included seven different behavioral choices; students were asked to rate every choice on a scale ranging from 1 (*not a very good choice*) to 7 (*a very good choice*). Both versions shared 15 identical items and differed in 10 items. Thus, each student received 25 vignettes. These items were scored in two different ways so that the scores assigned to each participant were derived by calculating the squared Euclidean distance (d^2) and the Mahalanobis distance (D^2; the standardized Euclidean distance, where not only absolute distances but also their dispersions are taken into account) of the participant's ratings for each possible solution strategy from the mean ratings of the total sample.

For each of the seven possible solution strategies accompanying each of the 25 vignettes, the sample's mean rating was subtracted from the participant's rating. These computations resulted in a vector of seven simple difference scores for each participant for each of the 25 vignettes presented in one of the versions, and thus $25 \times N$ vectors in all. For the Euclidean distance, the "ordinary" distance between the seven points of the individual's profile and the mean profile of the sample were calculated. For the Mahalanobis distance, the vectors of different scores were each multiplied by the variance–covariance matrix of the seven possible response strategies from which the difference scores were created. The resulting 7×1 vector was then multiplied by the transpose of the original difference-score vector, resulting in a scalar, D^2. These computations, then, resulted in 25 d^2 and D^2 values per individual, one per vignette, and thus $25 \times N$ in all. The individual's total score for the School Life Inventory was determined by averaging the resulting vignette-level values (25 vignettes for each participant).

Although there is a conceptual difference between using the Euclidean distance measure (d^2) and the Mahalanobis distance measure (D^2), where the former operates only with absolute differences and the latter also includes variation in these distances, the Pearson correlations between d^2 and D^2 were .97, .97, and .96 for Times 1–3, respectively. As noted earlier, scores on the practical ability performance measures were determined with reference to the average, or consensus responses of the sample. Important concerns arise when consensual scoring techniques become imbalanced with regard to race, ethnicity, or gender, as such imbalances might be biased against minority group members; other problems arise with regard to defending the basis of any particular individual's score against the average responses of the sample. However, using an "expert group" as a reference instead of the average responses of the sample might lead to similar problems, for example, with determining the demographic characteristics of those individuals comprising such an expert group. Legree (1995) demonstrated that the ratings of experts and nonexperts on a situational judgment inventory are highly correlated, indicating that a fairly knowledgeable nonexpert consensus can be as sensitive to relative differences in solution quality as an expert consensus. Mayer and colleagues (Mayer, Salovey, Caruso, & Sitarenios, 2003) have shown that an expert panel shows more within-group consistency than does a general sample in selecting the "correct" answer on emotional intelligence items; however, there appears to be a great

deal of between-group agreement in terms of these items, suggesting that both expert panels and general samples tend to agree on the overall correct answers to emotional intelligence items. For the two versions, Cronbach's αs for d^2 were .91 and .93 (Time 1), .95 and .95 (Time 2), and .94 and .94 (Time 3). For D^2, Cronbach's αs were .84 and .83 (Time 1), .92 and .95 (Time 2), and .90 and .93 (Time 3). Given the similarity of the scores obtained by these scoring systems, only the results obtained by calculating the Mahalanobis distance (D^2) are presented here, where not only absolute distances between an individual's and the mean profile of choice are taken into account but also the corresponding variance. Notice that in this task, the higher the score, the greater the deviation from the mean profile of answers, and thus the worse the score.

The practical reasoning in writing assessment contained eight prompts presenting different everyday situations comparable to those experienced by the school's students. For example, students could be presented with the situation of having an assignment due the next day when they had forgotten about the assignment and the deadline. Participants were asked to continue the story by identifying with the main character and developing the next step in the plot. The written products were subsequently reviewed by two independent raters (the same pair of raters for Times 1 and 2 and a different pair—with one continuous rater and one new rater—for Time 3) using scoring rubrics. The corresponding median κs between the two raters were .95, .96, and .93 for Times 1, 2, and 3, respectively.

The ratings were done on three dimensions per item: (a) quality of writing (grammar, spelling, and so on); (b) practicality of the proposed solution; and (c) quality of argument delivered in the excerpt. To process the ratings, the data were subjected to Rasch analysis using Facets (2009); three combined scores per participant were generated per time wave. The reliability estimates for these scores were .83, .61, and .72 (Time 1); .70, .64, and .74 (Time 2); and .73, .66, and .79 (Time 3) for quality of writing, practicality, quality of argument, respectively. These reliability indices indicate the consistency with which the same group of people will score on a different set of items of comparable difficulty and discrimination. The median (across quality of writing, practicality, and quality of argument) reliability estimates for the items (i.e., consistency of responses across the eight practical reasoning tasks) across the three waves of data collection were .96, .95, and .76 for Times 1–3, respectively. These indices suggest that these tasks would behave with similar consistency as a scale when administered to a different group of students.

The creative writing task asked for a brief story under one of five proposed titles: (a) "Too Much, Too Fast"; (b) "The Landing on the Planet Vespa"; (c) "Third Time's the Charm"; (d) "The Spy Was Not Captured After All"; and (e) "When the Music Stopped." Across the three waves, Titles 1–4 were chosen with almost equal frequency (20%–25%); Title 5 was chosen by 8%–11% of participants. Only 4 participants chose the same titles over the three data collection waves (Titles 2 and 5, 2 participants for each title), but a number of participants selected the same title twice (11, 2, 6, 2, and 8 selections for Titles 1–5, respectively).

A number of participants provided answers that were unscorable, that is, that did not correspond to the title/instruction or were not legible (*n*s = 7, 11, and 16 for Times 1–3, respectively). Only 1 participant did not generate ratable answers for any of the three waves. Two independent evaluators, using a scale of 1–6 (low to high), rated the story on two dimensions: (a) quality of written expression and (b) creativity. The corresponding average κs were .33, .72, and .82 for Times 1–3, respectively.

To process the ratings, the data were subjected to Rasch analysis using Facets; a combined score (a synthesis of quality of written expression and creativity) was generated per time wave. Because five different titles were used in this task, three facets were introduced to score these data: title, rater, and participant. The participant-based reliability estimates for these combined scores were .89, .86, and .84 for Times 1–3, respectively. These reliability indices indicate the consistency with which the same group of people will score on a different set of items of comparable difficulty and discrimination. The reliability estimates for quality of written expression and creativity were .98, .88, and .69 for Times 1–3, respectively. Similarly, the five different titles showed acceptable levels of reliability: .71, .83, and .85 for Times 1–3, respectively, indicating that these (or a comparable set of) titles have the potential to generate reliable data when administered to a different sample of participants.

The scientific reasoning task included 15 different word problems describing various situations related to the application of scientific knowledge; students were asked to find a solution using some knowledge of the sciences. Answers were subsequently reviewed, using scoring rubrics, by two independent raters (the same pair) for Times 1 and 2 and by three independent raters (all but one different from Times 1 and 2) for Time 3. Each item was scored on a 4-point scale ranging from 0 (*completely wrong*) to 4 (*completely right*). The corresponding median κs were .92, .94, and .85 (calculated across the three raters) for Times 1, 2, and 3, respectively. The raters' ratings were subjected to Rasch analysis with Facets. For this scale, only one score per participant was acquired, and corresponding rater reliability estimates were .57, .76, and .85 for Times 1–3, respectively. These reliability indices indicate the consistency with which the same group of people will score on a different set of items of comparable difficulty and discrimination. The reliability estimates for the items (i.e., consistency of responses across the 15 scientific reasoning tasks) across the three waves of data collection were .96, .96, and .97 for Times 1–3, respectively. These indices suggest that these tasks would behave with similar consistency as a scale when administered to a different group of students.

The self vs. other(s) attribution of blame assessment included eight brief stories with six possible answer choices. Students were expected to circle only one answer per story, with each choice having a specific expert-determined weight: 0, 1, or 2. Higher values indicate the more expert-valued choices of answers. These scores were generated using a rubric developed by three professional psychologists based on modern literature on moral reasoning. The rubric was then used in scoring by two raters, who were trained prior to scoring to achieve a consensus for at least 80% of the ratings. The scale assessed the student's capacity to determine whether a situation is attributable to his or her own doing, could be someone else's fault, or is viewed as a result of "fate." Cronbach's αs were, disappointingly, only .25, .31, and .48 for Times 1, 2, and 3, respectively.

Teacher reports. Teachers were asked to rate students, using a scale from 1 to 7, on 23 items subgrouped into three categories: self-perception and coping skills (SP&CS, seven items), academic

skills (AS, eight items), and social and practical skills (S&PS, eight items). Here are some illustrative items: "This student can cope with failure"; "Academic skills of this student are strong"; "This student has good relationships with peers," for the three categories, respectively. Twenty teachers from two academic domains—7 teaching various levels of English and 13 teaching various levels of mathematics—were asked to provide five sets of evaluations spread out evenly over time from late November to the end of the students' first year at Choate. Each student was evaluated by two teachers, one in English and the other in math. Teachers were not asked for evaluations earlier in the academic year because they felt that such evaluations would be inaccurate or, at best, less accurate, as they did not know their students as well. In these evaluations, a reversed Likert scale was used, with 1 indicating the strongest presence of a positive characteristic. Because there were so many different raters involved for the overall group, and two raters per student, the data were processed with Facets; three continuous indicators were generated (SP&CS, AS, and S&PS) for all five waves of evaluation. Table 4 summarizes the reliabilities for students, raters, and items for Times 1–5.

School reports. The school provided (a) demographic data (gender, age, ethnicity); (b) grades for all subjects the students took during their first and second years at Choate, summarized as GPAs for the fall, winter, and spring trimesters for 2 consecutive years; (c) information on financial aid (whether the student received any aid or not); and (d) preadmission data consisting of students' GPAs in their previous middle schools and their standardized test results (i.e., three subscores: SSAT Verbal, Quantitative, and Reading).

Results

Findings from this study are presented in three blocks. First, descriptive statistics for all waves of data collection are presented. Then, changes over the first year of education at the school across multiple assessment times are considered in both student and teacher reports. In addition, the changes in Choate GPA over six trimesters are considered. Finally, the power of indicators collected prior to admission by the school and while at the school as predictors of academic achievement (i.e., average Choate GPA) is considered.

Descriptive Statistics

Table 5 presents the descriptive statistics for the various indicators used in this study. Similar to the procedures in the Pilot Study, we screened the outcome variables of interest (GPAs from six trimesters and the GPA averaged across the six trimesters) for group differences on the basis of gender and ethnicity status variables. None of the outcome variables demonstrated group differences, so we carried out all subsequent analyses for the whole sample without stratification by gender or ethnicity.

Changes Over the First Year in Independent Variables and Over the Six Trimesters in Choate GPA

The next set of analyses addresses the question posed by the SRL perspective: We wanted to investigate whether we could detect any changes in the indicators of Choate GPA and competencies in SRL across the first year in the new schooling environment. Specifically, were there any fluctuations in these students' GPAs over their time at Choate? Also, were there any changes across the full first year at the new school in the students' self-reports and performance indicators, and the teachers' evaluations of their students? And, finally, were the changes (if any) in the independent variables related to the changes in Choate GPA?

To analyze these changes, using HLM 6.0, we first fitted an unconditional linear growth model for Choate GPA and all independent variables measured multiple times. We started by fitting an unconditional linear growth model using the GPAs over six trimesters (freshman and sophomore years) at Choate. The results fitted the pattern of GPAs apparent from Table 5: There was linear growth in GPAs over time, with the starting point at approximately 3.14 and a growth rate of .014 GPA per month or .040 GPA per trimester, $T(871) = 8.78$, $p < .001$. These results are similar to those obtained in the Pilot Study, although both the starting point and the average growth of GPAs in this Choate class as a whole appears to be greater compared with these parameters among the three Icahn classes. Given the goal of the Icahn Scholars Program to accept students to Choate who would be at risk of academic failure were they to stay home, this result seems appropriate.

We carried out similar analyses for all PACE indicators on which we had multiple measurements; the *time scale* was defined as months at Choate.

Table 4
Reliability Indicators of Teacher Surveys (Study 1, Main Study)

Indicator	Reliability	Evaluation waves				
		Time 1	Time 2	Time 3	Time 4	Time 5
Self-perception & coping skills	Students	.87	.88	.89	.91	.90
	Raters	.91	.86	.95	.92	.95
	Items	.93	.87	.80	.81	.84
Academic skills	Students	.90	.91	.89	.95	.92
	Raters	.95	.94	.95	.95	.95
	Items	.98	.97	.80	.96	.95
Social & practical skills	Students	.75	.81	.83	.86	.77
	Raters	.91	.94	.93	.93	.94
	Items	.97	.98	.95	.95	.98

Table 5
Means and Standard Deviations for Indices Used With Choate Students in Study 1, Main Study

Assessment	December		January		March		April		May	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Students' self-report indicators										
Academic Self-Efficacy	7.00	0.90	6.86	0.99			6.99	1.04		
Academic Locus of Control	3.53	0.47	3.49	0.57			3.47	0.51		
Academic Motivation	3.86	0.54	3.89	0.64			3.86	0.63		
Students' performance indicators										
School Life Inventory	6.92	3.04	6.97	3.16			6.96	2.90		
Practical reasoning in writing										
Quality of writing	−0.22	1.88	−0.75	1.75			−0.57	1.77		
Practicality	0.00	0.84	0.00	0.88			0.00	0.90		
Quality of argument	0.00	1.07	0.00	1.13			0.00	1.47		
Creative writing	−0.93	3.76	−2.44	4.44			−5.60	5.44		
Scientific reasoning	0.00	0.51	0.00	0.63			0.00	0.81		
Attribution of blame	0.99	0.30	1.09	0.31			0.97	0.37		
Teacher survey[a]										
Self-Perception & Coping	−1.05	0.91	−1.29	1.03	−1.63	1.40	−2.18	1.40	−2.13	1.46
Academic Skills	−1.62	1.00	−1.84	1.15	−1.92	1.78	−2.18	1.75	−2.21	1.74
Social & Practical Skills	−0.99	0.62	−1.73	0.78	−1.57	1.24	−2.02	1.10	−2.28	1.12
GPA										
Pre-Choate	3.78	0.33								
	Fall				Winter				Spring	
Choate										
Freshman year	3.20	0.43			3.20	0.51			3.20	0.52
Sophomore year	3.39	0.56			3.40	0.62			3.37	0.61
SSAT										
Verbal	320.49	14.18								
Quantitative	325.99	13.88								
Reading	313.53	13.89								

Note. GPA = grade point average; SSAT = Secondary School Admission Test.
[a] The scale was reversed, with 1 indicating the highest and 7 indicating the lowest rating.

Overall, there was little change in students' indicators of SRL. Specifically, the unconditional model showed no growth in Academic Self-Efficacy across the three time points. The unconditional model showed a tendency for decline in the internal Academic Locus of Control across the three time points ($B = -0.013$, $SE = 0.01$), $T(437) = -1.74$, $p < .1$, indicating that, with the passing of time, students' orientation toward academics tended to diminish, although this change was not statistically significant. There was no change detected in Academic Motivation over time with an unconditional growth model.

Similarly, we detected little or no change in students' performance measures. The unconditional model of the School Life Inventory detected no change in the students' tacit knowledge of the school's environment across the period from December to May of the freshman year. The practical reasoning task generated three indicators, as described earlier: (a) quality of writing (grammar, spelling, and so on); (b) practicality of the proposed solution; and (c) quality of argument delivered in the excerpt. Correspondingly, three different linear growth models were fitted in these data. The unconditional model fitted into the data for quality of writing indicated a decrease in this indicator over time at the rate of about .08 per Choate academic month ($B = -0.08$, $SE = .03$), $T(437) = -2.52$, $p < .01$. The unconditional model for the indicator of practicality did not show any change in performance over time. Fitting the unconditional model into the data for the quality of argument indicated that there was no change over three assessment times in this variable either. The creative writing unconditional model indicated a substantial change in performance over the three waves of data collection, with the quality of writing getting worse at approximately .8 of the unit of judgment per Choate academic month ($B = -0.81$, $SE = 0.09$), $T(437) = -9.34$, $p < .001$. The unconditional model of scientific reasoning and attribution of blame did not register any changes over time in the performance on these tasks. Further research is needed to understand the mechanisms of these changes (or lack of such); they may be connected to deep motivational structures or to simple exhaustion from accepting and managing repeated tasks over time.

This set of analyses was completed with growth modeling of teacher surveys. As indicated earlier, teacher surveys were scored to generate the following indicators: (a) SP&CS, (b) AS,

(c) and S&PS. The scoring was reversed, with 1 indicating the highest and 7 indicating the lowest rating. Unconditional models fitted for the ratings of students' skills captured a substantial change in these ratings over time in all three domains of teacher ratings ($B = -0.23$, $SE = 0.01$), $T(747) = -17.69$, $p < .001$; ($B = -0.11$, $SE = 0.01$), $T(747) = -8.18$, $p < .001$; ($B = -0.21$, $SE = 0.01$), $T(747) = -16.61$, $p < .001$), for SP&CS, AS, and S&PS, respectively. These findings suggest that teachers' perceptions of students' competencies increased substantially in value over the freshman year. This increase appears to happen at about .2 of the unit of judgment per Choate academic month.

Summarizing these results, three observations need to be made. First, similar to the results in the Pilot Study, Choate GPA showed a significant fluctuation depending on the period of time spent at the school. Second, only one of the self-report scales, Academic Locus of Control, showed fluctuations over time. This result is somewhat contrary to expectations, based on the SRL perspective, and deserves further investigation. Third, all dimensions of the teacher ratings improved with time; thus, it appears that teachers increase their appreciation of the school's students as they get to know them better.

These data provide us with a rather unique opportunity to consider the time-based dynamics in Choate GPA in conjunction with time-based dynamics in the PACE Battery indicators and teacher ratings. Using the software Mplus Version 3 (Múthen & Múthen, 2005), we carried out a set of analyses in which two processes were considered simultaneously, which estimated the associations between the intercepts and slopes of these two processes.

When we analyzed changes in Choate GPAs in conjunction with fluctuations in the indicators from the PACE Battery, a number of significant findings were established; the mean values of Choate GPA were associated with both SRL self-ratings and performance tasks. Specifically, higher mean Choate GPA was associated with higher levels of Self-Efficacy ($B = 0.10$, $SE = 0.04$; $T = 2.68$, $p < .01$), and higher levels of Academic Motivation ($B = 0.05$, $SE = 0.02$; $T = 2.74$, $p < .01$). Similarly, higher Choate GPAs were associated with a number of performance task indicators: better tacit knowledge of Choate ($B = -0.37$, $SE = 0.11$; $T = -3.38$, $p < .001$, for the School Life Inventory); higher performance on the practical reasoning task ($B = 0.11$, $SE = 0.06$; $T = 1.88$, $p < .05$, for quality of writing); higher creativity in writing ($B = 0.36$, $SE = 0.14$, $T = 2.50$, $p < .05$, for creative writing); higher levels of scientific reasoning ($B = 0.06$, $SE = 0.02$, $T = 2.84$, $p < .05$); and more adequate attribution of blame ($B = 0.03$, $SE = 0.01$, $T = 2.85$, $p < .01$).

Only two estimates of the rate of change of the indicators from the PACE Battery were associated with the mean Choate GPA: the Practicality aspect of the practical reasoning task ($B = 0.07$), $SE = 0.03$, $T = 2.16$, $p < .05$), and the scientific reasoning task ($B = 0.07$, $SE = 0.03$, $T = 2.09$, $p < .05$). None of the indicators from the PACE Battery predicted the rate of growth in Choate GPA.

The facets of teacher ratings also showed numerous associations with time fluctuations in Choate GPA (see Table 6 for details). Specifically, dynamics of teacher evaluations of the students on all three dimensions suggest an association of the mean value of Choate GPA (the higher the GPA, the higher the evaluations).

Table 6
Parallel Changes in Teacher Ratings and Choate GPA (Study 1, Main Study)

Teacher ratings	Choate GPA	
	Intercept (B, T)	Slope (B, T)
Self-Perception and Coping Skills		
Intercept	−.11, −3.12**	
Slope	−.26, −7.44***	−.001, −2.42*
Academic Skills		
Intercept	−.21, −5.66***	
Slope	−.32, −8.51***	−.001, −2.41*
Social and Practical Skills		
Intercept	−.07, −1.97*	
Slope	−.17, −4.90**	

Note. GPA = grade point average; B(s) indicate estimated parameters; T(s) indicate corresponding T ratios.
* $p < .05$. ** $p < .01$. *** $p < .001$.

In addition, it is apparent that the mean value of Choate GPA was reflected by the rate of growth in the teacher evaluations (for all three variables: SP&CS, AS, and S&PC). Finally, the rate of growth of Choate GPA is associated with the rate of growth in SP&CS and AS. In summary, these results suggest that changes in Choate GPA and changes in teacher ratings are parallel and that it is the dynamics of GPA that are associated with the ratings (i.e., teacher perception of students changing depending on how successful they are at Choate).

The results of these analyses allowed us to make two general observations. First, the mean GPA at Choate appears to be related to a number of psychological constructs captured by the PACE Battery. However, the causal links between these associations are difficult to establish at this point, as a result of the design of the study, the relatively low power of our sample, and, possibly, the influences of other factors. Second, the mean GPA and teacher ratings also appear to be associated; in fact, these associations are strong, and their patterns allow us to hypothesize the causal impact of GPAs on teachers' perceptions of their students.

Predicting First-Year GPA On the Basis of Pre- and In-School Assessments

Up to this point, we have shown that, in the Main Study, (a) we collected reliable data on the newly developed indicators, and (b) these indicators are relevant to predicting the mean values of Choate GPA. Thus, we have already demonstrated the value of the PACE Battery in capturing the average value of GPA at Choate. What we have not yet demonstrated is the relative value of the PACE indicators when compared with the predictive validity of pre-Choate GPA and the three SSAT indicators.

In this section of analyses, we address the specific question of whether students' GPA(s) across the six trimesters of their freshman and sophomore years at Choate can be predicted incrementally, in terms of their absolute value and rate of change, by any of the newly developed measures over the indicators of SSAT and pre-Choate GPA.

To address these questions, we performed two sets of analyses. First, we preserved the time-based variability in the GPAs but constrained the variability in independent measures. In other words, we obtained the first principal components of all student- and teacher-generated indicators across time and ran a set of analyses similar to those described earlier: We let GPAs vary in time, but our predictors were all single-point indicators.

Second, we obtained an average GPA across the six terms and completed a set of traditional hierarchical linear regression analyses, evaluating increments in prediction over SSAT indicators and pre-Choate GPA.

Conditional Growth in Choate GPA: What Matters?

Conceptually, these analyses are similar to those performed in Study 1, but in this study, the PACE Battery contained more indicators.

The summary of these analyses is shown in Table 7. When single-predictor models were fitted, 13 of the 14 investigated indicators were found to be associated either with the intercept or with the slope of the Choate GPA growth model. When considered in a single model, however, a number of indicators appeared redundant. To lessen the redundancy of the information, the reduced model was fitted and the reduction of the number of the parameters did not worsen the model fit (χ^2 difference with 10 *df* was 13.95, *ns*). The final reduced model indicated that the Choate GPA intercept is predicted by pre-Choate GPA, SSAT Quantitative, and two PACE Battery indicators—Academic Self-Efficacy and School Life Inventory. The rate of growth of GPA was predicted by pre-Choate GPA, SSAT Quantitative, and Academic Motivation.

To summarize, these analyses attempted to differentiate growth profiles of Choate GPA and succeeded in doing so with both traditional (i.e., pre-Choate GPA and standardized tests) and novel (WICS- and SRL-based) measures. The remaining question, however, addresses the relative predictive validity of WICS- and SRL-based measures over and above the pre-Choate GPA and standardized tests.

Predicting Average (Freshman and Sophomore) Choate GPA

As we indicated earlier, the school provided preadmission data for this class of freshman. Specifically, SSAT indicators and pre-Choate GPAs were available for these comparative analyses. However, these indicators were available only for a single time point. Correspondingly, all other indicators were condensed from multiple points of measurement into summative indicators by means of principal-component analyses.[8]

We carried out two sets of analyses. In the first set, we applied a hierarchical regression analysis, in which the three SSAT indicators were entered first in a single step and then followed by (a) indicators of practical competence (Practical Reasoning and School Life Inventory), (b) indicators of creative competence (creative writing and scientific reasoning), (c) the indicator of ethical reasoning (as a partial measure of wisdom), and (d) the SRL-related indicators (Academic Self-Efficacy, Academic Locus of Control, and Academic Motivation). The second set was identical to the first set with one difference: SSAT indicators were entered simultaneously with pre-Choate GPA. The results of both sets of analyses are shown in Table 8.

The top portion of Table 8 shows that, overall, the introduction of WICS- and SRL-related indicators increased the amount of explained variance in the average Choate GPA by approximately 165%. Specifically, the amount of variance explained by all the indicators of SSAT was approximately14%, and the introduction of indicators only of practical competence almost doubled this amount (27.2%). Subsequently, both indicators of creative competence and ethical reasoning contributed to the explained variance, with self-report indicators bringing the values of R^2 to 37%. Much like the results of growth modeling, indicators that contributed the most to this profile were School Life Inventory and Academic Self-Efficacy.

The bottom portion of Table 8 repeats the first set of analyses but includes Pre-Choate GPA as a predictor at the first step. It appears that middle-school GPA is the best predictor across the board, explaining about29% of freshman and sophomore GPA. As is often the case, the best predictor of later GPA is earlier GPA. After SSAT and middle-school GPA, the WICS- and SRL-related measures together explain about 10% of the variance. In other words, they explain approximately as much as all three indicators of SSAT do when they are presented collectively but with no other predictors in the regression equation. When SSAT indicators are entered after middle-school GPA, they explain only an additional 8% of the variance, that is, less than what is explained by the PACE Battery measures with SSAT indicators and pre-Choate GPA in the regression equation.

In summary, the WICS- and SRL-related measures substantially increase predictive power for Choate GPA over SSAT alone, given, of course, that SSAT was used in the admission of students and thus was restricted in range.

Discussion

In this study, we generally confirmed and expanded the findings from our Pilot Study. Specifically, we showed, on a larger, more generalizable sample, that measures derived from the WICS and SRL theoretical frameworks substantially increased the predictability of Choate GPA. In addition, in conjunction with traditional measures of pre-Choate GPA and standardized tests, the measures of the PACE Battery predict not only the mean value of GPA but also its rate of growth. The magnitude of predictive validity is substantial, totaling up to 50% of the variance in Choate GPA. Although the best single predictor of Choate GPA appears to be middle-school GPA, measures from the PACE Battery predict as much as or more variance than do indicators of SSAT. They also contribute substantially, independent of SSAT.

---

[8] These are indicators of cumulative percentages of variance explained by the first principal component across multiple time measures: (a) Academic Self-Efficacy—63.7%; (b) Academic Locus of Control—71.4%; (c) Academic Motivation—75.2%; (d) School Life Inventory—74.4%; (e) practical reasoning task: quality of writing—63.7%, practicality—74.0%, quality of argument—56.2%; (f) creative writing—52.5%; (g) scientific reasoning—61.5%; (h) attribution of blame (58.7%).

Table 7
Conditional Change in Choate GPA Over Freshman and Sophomore Years at Choate (Study 1, Main Study)

Model	Intercept (starting value)			Slope (growth rate)		
	Coefficient	*T*	*p*	Coefficient	*T*	*p*
			Single predictor			
Parameter model						
Pre-Choate GPA	.9146	8.93	<.001	.0241	4.88	<.001
SSAT Verbal	.0078	2.77	.006	.0000	0.18	.855
SSAT Quantitative	.0130	4.71	<.001	.0006	5.07	<.001
SSAT Reading	.0070	2.23	.015	−.0001	−0.53	.595
Academic Self-Efficacy	.1374	3.49	.001	−.0012	−0.82	.415
Academic Locus of Control	.0704	1.74	.082	.0005	0.33	.744
Academic Motivation	.1532	3.93	.039	.0053	3.46	.001
School Life Inventory	−.1490	−3.79	<.001	−.0014	−0.88	.378
Practical reasoning						
Quality of writing	.0900	2.23	.026	.0015	0.97	.334
Practicality	.1308	3.31	.001	.0034	2.26	.024
Quality of argument	.0441	1.08	.280	−.0007	−0.48	.628
Creative writing	.1169	2.94	.004	.0011	0.69	.487
Scientific reasoning	.1673	4.32	<.001	.0015	1.01	.313
Attribution of blame	.1082	2.71	.007	−.0005	−0.34	.736
			Multiple predictor (combined)			
Full model						
Pre-Choate GPA	.6762	7.13	<.001	.0156	3.00	.003
SSAT Verbal	.0014	0.54	.587			
SSAT Quantitative	.0101	4.41	<.001	.0005	4.47	<.001
SSAT Reading	.0017	0.72	.471			
Academic Self-Efficacy	.0749	2.23	.025			
Academic Locus of Control	.0017	0.53	.958			
Academic Motivation	.0170	0.52	.605	.0028	1.69	.090
School Life Inventory	−.0780	−2.28	.022			
Practical reasoning						
Quality of writing	.0289	0.93	.351			
Practicality	.0303	0.93	.357	.0026	1.62	.105
Creative writing	.0170	0.52	.605			
Scientific reasoning	.0396	1.15	.251			
Attribution of blame	.0098	0.31	.759			
Reduced model						
Pre-Choate GPA	.7355	7.88	<.001	.0171	3.34	.001
SSAT Quantitative	.0105	4.84	<.001	.0005	4.22	<.001
Academic Self-Efficacy	.0978	3.28	.001			
Academic Motivation				.0036	2.31	.021
School Life Inventory	−.1151	−3.86	<.001			

Note. GPA = grade point average; SSAT = Secondary School Admission Test.

Study 2

On the basis of the results from Study 1, we made the decision to introduce selected measures from the PACE Battery into Choate's admission process. To minimize the associated costs and to avoid the need for proctoring, we included only self-assessments; they were presented to applicants as an "optional addition" to their application packages. In addition, during candidates' evaluations, the Choate admissions office representatives rated their perceptions of the creative and practical abilities of the interviewees, based on the interviewees' answers in the admission application to two short essay questions that were developed by Choate staff in concordance with the WICS theory to elicit a creative or practical response; the representatives used a scale ranging from 1 (*high*) to 3 (*low*), mimicking, at low cost, the performance measures used in Study 1. Like many measures in the admission process, this tool was used in an effort to either distinguish a candidate, positively or negatively, or to corroborate with other evidence in the application, whether it be the interview, other writing samples or the confidential teacher recommendations. All participation in this research was voluntary.

Method

The main objective of this study was to investigate the power to predict first-trimester Choate GPA on the basis of a number of conventional assessments (i.e., previous school GPA and SSAT) and augmenting assessments (i.e., self-report instruments from the PACE Battery).

Participants

For the academic year in which the study was conducted, Choate had 1,495 applicants (n = 685, or 45.8% female and n =

Table 8
Incremental Prediction of High School GPA Using WICS- and SRL-Related Measures (A) Above and Beyond SSAT and (B) Above and Beyond SSAT and Pre-Choate GPA (Study 1, Main Study)

Measure	Step 1	Step 2	Step 3	Step 4	Step 5
			A		
SSAT					
Verbal	.097	.052	−.009	−.011	.050
Quantitative	.304***	.374***	.398***	.397***	.386***
Reading	.071	.011	−.004	−.003	−.031
PACE WICS Indicators					
Practical competence					
Practical reasoning		.208**	.127	.125	.107
School Life Inventory		−.267***	−.209**	−.197*	−.191*
Creative competence					
Creative writing			.137	.134	.085
Scientific reasoning			.173*	.172*	.137
Ethical reasoning					
Attribution of blame				.031	.022
PACE SRL indicators					
Academic Self-Efficacy					.176*
Academic Locus of Control					−.012
Academic Motivation					.120
R^2	.142	.272	.318	.319	.374
			B		
SSAT					
Verbal	.057	.032	−.002	.004	.043
Quantitative	.193**	.250***	.270***	.269**	.271***
Reading	.136	.088	.075	.076	.048
Pre-Choate GPA	.552***	.494***	.468***	.467***	.440***
PACE WICS Indicators					
Practical competence					
Practical reasoning		.120	.078	.076	.072
School Life Inventory		−.201**	−.168**	−.158*	−.156*
Creative competence					
Creative writing			.072	.070	.038
Scientific reasoning			.109	.107	.083
Ethical reasoning					
Attribution of blame				.026	.019
PACE SRL indicators					
Academic Self-Efficacy					.147***
Academic Locus of Control					.010
Academic Motivation					.054
R^2	.433	.491	.506	.506	.535

Note. n = 152. Entries are standardized beta coefficients. GPA = grade point average; WICS = Wisdom, Intelligence, Creativity Synthesized; SRL = self-regulated learning; SSAT = Secondary School Admission Test; PACE = Psychology of Abilities, Competencies, and Expertise.
* $p < .05$. ** $p < .01$. *** $p < .001$.

810, or 54.2% male). Of these applicants, 377 (25.2% of the total sample, or 186 [or 49.3%] females and 191 [or 50.7%] of males) engaged the option of completing self-report assessments: the Academic Self-Efficacy Scale, the Academic Locus of Control Scale, and the Intrinsic/Extrinsic Academic Motivation Scale. The applicants who took the self-assessments did not differ from the rest of the sample on gender and age or SSAT Quantitative. Yet, those who took additional assessments tended to have slightly higher GPA at their previous school (3.67 vs. 3.55, $p < .001$), and slightly higher Reading (689.76 vs. 679.79, $p < .05$) and Verbal (705.13 vs. 694.43, $p < .01$) SSAT scores. Also, compared with the ethnic profile of the total pool of applicants, White students were overrepresented among those students who took the self-assessments (56% vs. 38.6% among those who took and did not take the assessment, respectively).

The school admitted 259 students (115 girls and 144 boys, mean age = 173.6 months, SD = 16.06),[9] 88 of whom took the self-assessments. These 88 students were predominantly White (61.7%), but many students (38.3%) came from a variety of other ethnic groups. Among the admitted students who took the assessments, girls were slightly overrepresented (53.4% vs. 39.8% among those who were admitted and did not take the assessment, respectively), but there were no ethnicity-based differences. In

[9] These students were admitted into a variety of classes and grades.

addition, although there were no SSAT-based differences among students with and without self-assessments, the GPA differences remained (3.92 vs. 3.62, $p < .001$).

Procedure

All self-reports were completed online. The applicant was requested to log in to a secure Web site, where the questions were presented electronically, the answers were recorded on a remote server, and the data were scored and delivered to Choate's admissions office. The admission decisions were made irrespective of the SRL indicators.

Measures

Self-reports. Applicants completed the Academic Self-Efficacy, Academic Locus of Control, and Intrinsic/Extrinsic Academic Motivation scales (see detailed descriptions in Study, Main Study). In this study, for the total sample ($N = 377$), Cronbach's αs were .80, .73, and .72 for the three scales, respectively. These alphas are comparable to those obtained in the Main Study of Study 1.

School reports. The school provided (a) demographic data (gender, age, ethnicity, and citizenship), (b) grades for all subjects the students took during their first trimester at Choate, summarized as fall GPA, and (c) preadmission data consisting of students' GPAs in their previous middle schools and their standardized test scores (i.e., SSAT Verbal, Quantitative, and Reading).

As mentioned earlier, admissions officers at Choate ranked each applicant on two dimensions, Creativity and Practicality. Each applicant was rated by two admissions officers. Their ratings were summarized by principal-component analyses, and the first component scores were saved for subsequent analyses. The first components explained 72% and 71% of the variance for Creativity and Practicality, respectively.

Results

Descriptive Statistics

Table 9 presents the descriptive statistics for the variables of interest (the PACE Battery indicators and the admissions officers' ratings, the SSAT indicators, and pre-Choate GPA) estimated for the whole sample of applicants and for the group of students admitted to Choate. First-trimester Choate GPA was, clearly, available only for the students ultimately admitted to Choate. As is apparent from Table 9, students admitted to Choate differed from nonadmitted students on pre-Choate GPA, $F(1, 1326) = 16.56$, $p < .001$, with the admitted students having higher GPA; SSAT Verbal, $F(1, 1293) = 10.00$, $p < .01$, with the admitted students having higher SSAT Verbal; and ratings of Creativity, $F(1, 1114) = 7.74$, $p < .01$, with the admitted students having lower ratings of creativity. There were no differences among admitted and nonadmitted students on the PACE SRL indicators; but again, because the completion of these self-reports was voluntary and the admission decisions were made irrespective of these indicators, the lack of group difference should be interpreted with caution, because possible biases cannot be ruled out.

When the demographic variables of gender and ethnicity status were considered, there were no group differences in the outcome variable (first-trimester Choate GPA). Thus, neither of these variables was included in subsequent analyses.

Predicting First-Trimester Choate GPA

Although the sample of volunteers who took self-report scales and were admitted to Choate was relatively small ($n = 88$), it provided us with an opportunity to investigate the predictive validity of pre-Choate GPA, the SSAT indicators, the PACE SRL indicators, and the admission ratings of Creativity and Practicality for first-trimester Choate GPA. Similar to the comparable regres-

Table 9
Means and Standard Deviations for Indices Used With Choate Applicants in Study 2

	Students who applied to Choate		Students admitted to Choate	
Sample	*M*	*SD*	*M*	*SD*
Indicator				
GPA				
Pre-Choate	3.58	0.48	3.69	0.41
Choate			3.28	0.42
SSAT				
Verbal	697.17	67.33	710.88	62.54
Quantitative	723.66	72.38	728.79	66.98
Reading	682.34	66.32	688.25	72.93
PACE SRL indicators				
Academic Self-Awareness	93.42	8.29	94.39	8.01
Academic Locus of Control	63.47	5.95	63.51	5.44
Academic Motivation	53.31	4.14	53.57	4.30
Admission rating				
Creativity	0.00	1.000	−0.17	0.98
Practicality	0.00	1.000	0.01	1.01

Note. GPA = grade point average; SSAT = Secondary School Admission Test; PACE = Psychology of Abilities, Competencies, and Expertise; SRL = self-regulated learning.

sion analyses conducted in Studies 1 and 2, we fitted two regression equations (see Table 10).

Three observations can be made on the basis of the results presented in Table 10. First, regression results here are similar to those from Study 1, presented in Table 3. Overall, the SSAT measures appear to explain only about 15% of Choate GPA, with SSAT Quantitative as the only variable making a statistically significant contribution to this prediction. This limitation may be in part a result of the restriction of range emanating from the use of SSAT scores in selection. The very brief self-reports forming the PACE SRL assessments contribute approximately 10% of unique predictive variance, and this contribution is sustainable with or without the presence of pre-Choate GPA. Second, in concert with earlier observations from both the Pilot Study and Main Study (Study 1), external Academic Locus of Control appears to be associated with higher Choate GPA. Third, in contrast to findings from the Main Study in Study 1, the indicators of Creativity and Practicality did not demonstrate predictive validity with regard to Choate GPA. This result underscores the importance of using stronger measures of these constructs, such as creative or practical maximum-performance tasks rather than self-reports and ratings by others, as has been done at the tertiary level (see Sternberg, 2007).

Table 10
Incremental Prediction of First-Trimester GPA Using SRL-Related Measures (A) Above and Beyond SSAT and (B) Above and Beyond SSAT and Pre-Choate GPA in Study 2

Measure	Step 1	Step 2
	A	
SSAT		
Verbal	−.190	−.175
Quantitative	.484**	.538**
Reading	.014	−.013
PACE SRL indicators		
Academic Self-Efficacy		−.008
Academic Locus of Control		−.330**
Academic Motivation		.100
Admission rating		
Creativity		−.048
Practicality		−.097
R^2	.150	.244
	B	
SSAT		
Verbal	−.186	−.167
Quantitative	.332*	.366*
Reading	.067	.042
Pre-Choate GPA	.419***	.419***
PACE SRL indicator		
Academic Self-Efficacy		.036
Academic Locus of Control		−.320**
Academic Motivation		.032
Admission rating		
Creativity		−.056
Practicality		−.114
R^2	.304	.402

Note. Entries are standardized beta coefficients. GPA = grade point average; SRL = self-regulated learning; SSAT = Secondary School Admission Test; PACE = Psychology of Abilities, Competencies, and Expertise.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Discussion

This study builds on the evidence obtained from both the Pilot and Main Studies (Study 1) by providing additional data suggesting that the predictive validity of middle-school GPA and standardized tests can be enhanced by the introduction of additional theory-based measures, such as self-reports, providing information on students' SRL. Although the gain in explained variance is relatively modest (i.e., about 10% of R^2), it is important to consider the costs, for both applicants and schools, in administering and basing their decisions on standardized tests, such as the SSAT, if three brief self-reports account for as much as approximately 67% of the variance accounted for by the SSAT (i.e., 10% vs. 15%). In short, although standardized tests provide important and valid information, it is crucial not to overstate their predictive power; after all, many other factors regulating students' learning at Choate, among them, self-efficacy and locus of control, contribute to the academic success of students in the highly competitive environment of college preparatory schools.

GENERAL DISCUSSION

The research described in this article shows that it is possible to construct an assessment, based on modern cognitive and motivational theories, that enhances the prediction of academic success over and above that of traditional tests and preentry GPA. The purpose of these assessments is not to replace but rather to supplement traditional tests. In SRL, learner attributes other than ability-achievement estimates are sought to augment the predictive power of prior grade and standardized tests. In WICS, analytical skills are theorized to be important for academic and life success, but other skills are assumed to matter as well. Thus, both analytical skills, whether assessed by the traditional measures for secondary school admissions or evaluated in some other reasonably objective way, and other skills need to be taken into account while predicting success in a secondary school.

An advantage of the PACE assessments is that they move educators and researchers beyond the drill-and-kill ritual that has come to pervade preparation for standardized tests. One could certainly develop skills that would enhance performance on our expanded assessments but, in doing so, would also be developing skills one needs for both school and life success.

Many independent schools are seeking new ways of improving their selection process, and even more importantly, schools are interested in changing the way they select to focus on the factors that matter most to the development of the person and the student process. Historically, it has been important to Choate to expand the traditional boarding school population. Rather than rely on "feeder" private schools for its student population, Choate has sought to expand the schools from which it accepts students. The broader and more diverse the sending school group, the more reflective of the general population and hence the broader the learning experience for the community. The research presented in this article might encourage other schools presently seeking to expand their cultural diversity to offer support similar to Choate's Icahn Program or to create other programs.

Another specific application of the findings presented in this article is toward counseling families with regard to their children's admission to Choate or a recommendation to look for a different

educational environment. When students flounder in the independent-school setting, it is upsetting for both the student and the faculty of the school; both feel disappointment and dissatisfaction. Our research suggests that it is possible to improve the prediction of academic success and, in some cases, avoid the wounds of student failure, at least in one school. We believe it may be worthwhile to try a similar approach in other schools. It might be possible to construct a battery that would work across many schools, or it might be better to customize batteries to individual schools. Whichever the case, the prediction of school success apparently can be enhanced by thinking more broadly about the skills that are measured at the time of application.

It is also important that, from a psychological point of view, our data show that modern psychological theories can improve the prediction of performance in an independent-school setting, beyond the prediction gained by a time-honored psychometric admissions test and entry GPA. Such customary measures provide some prediction. But our measures could potentially add to this prediction and show that some of the gap between the data desired and the data attained can be narrowed using measures based on modern psychological theories, such as the WICS (Sternberg, 2003), as well as modern theories of SRL (e.g., Boekaerts et al., 2000).

Although promising, these studies are characterized by a number of limitations. Specifically, the data presented here, although collected from many Choate students and applicants across a number of years, are still data collected from only one school. Thus, the degree of generalizability of these findings is difficult to judge before similar studies are carried out in other secondary schools. Another limitation of this work is that, although theoretically grounded, its theoretical scope is inclusive of only two modern theories of academic learning and success, SRL and WICS. Clearly, there are many more theoretical approaches in the literature that, when included in the conceptualization of the admission battery, may enhance its predictive power. Finally, though cohesive and consistent across all studies with regard to indicating the importance of predictors of academic success at Choate outside of GPA and conventional tests, the findings from Study 1 and Study 2 are slightly inconsistent. Although this inconsistency can be explained by the novelty of the measures, which could be strengthened, or the diversities of the samples including both enrolled and potential Choate students, clearly more research is needed to support and enhance the findings presented here.

Yet, we believe that this work is of interest, not only because it provides data supporting the importance of skills other than memory and analytical abilities for success in school but also because it expands the very limited research literature on admission and education in secondary schools. After virtually any study, there is always more to do, but this work, hopefully, formulates a convincing example for other secondary schools to experiment both with ways of carrying out their admission practices and diversifying their student populations.

References

Bandura, A. (1996). *Self-efficacy: The exercise of control.* New York: Freeman.

Beattie, S., Grise, P., & Algozzine, B. (1983). Effects of test modifications on the minimum competency performance of learning disabled students. *Learning Disability Quarterly, 6,* 75–77.

Boekaerts, M., Pintrich, P. R., & Zeidner, M. (Eds.). (2000). *Handbook of self-regulation.* San Diego, CA: Academic Press.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning test* (No. 2000–1). New York: College Entrance Examination Board.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development.* New York: Psychology Press.

Facets. (2009). Facets (Version 3.65.0) [Computer software]. Beaverton, OR: Winsteps.

Hedlund, J., Wilt, J. M., Nebel, K. R., Ashford, S. J., & Sternberg, R. J. (2006). Assessing practical intelligence in business school admissions: A supplement to the graduate management admissions test. *Learning and Individual Differences, 16,* 101–127.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor established with a Likert-based testing procedure. *Intelligence, 21,* 247–266.

Lupkowski-Shoplik, A. E., & Assoline, S. G. (1993). Identifying mathematically talented elementary students: Using the lower level of the SSAT. *Gifted Child Quarterly, 37,* 118–123.

Luthar, S. S., Cicchetti, D., & Becker, B. (2000). The construct of resilience: A critical evaluation and guidelines for future work. *Child Development, 71,* 543–562.

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion, 3,* 97–105.

Mills, C. J., & Barnett, L. B. (1992). The use of the Secondary School Admission Test (SSAT) to identify academically talented elementary school students. *Gifted Child Quarterly, 36,* 155–159.

Múthen, L. K., & Múthen, B. O. (2005). *Mplus: Statistical analysis with latent variables. User's guide.* Los Angeles, CA: Author.

Noble, J., & Sawyer, R. (2002). *Predicting different levels of academic success in college using high school GPA and ACT composite score* (No. 2002–4). Iowa City, IA: ACT.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89,* 187–207.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). San Diego, CA: Academic Press.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2000). *HLM 5.* Lincolnwood, IL: Scientific Software International.

Rotter, J. B. (1990). Internal versus external control of reinforcement: A case history of a variable. *American Psychologist, 45,* 489–493.

Schuerger, J. M., & Dizney, H. F. (1967). The validity for ninth grade achievement of the SSAT and other admission criteria at a private secondary school. *Educational & Psychological Measurement, 27,* 433–438.

Schunk, D. H. (2005). Self-regulated learning: The educational legacy of Paul R. Pintrich. *Educational Psychologist, 40,* 85–94.

Sternberg, R. J. (1996). *Successful intelligence.* New York: Simon & Schuster.

Sternberg, R. J. (2003). *Wisdom, intelligence, and creativity synthesized.* New York: Cambridge University Press.

Sternberg, R. J. (2007). Finding students who are wise, practical, and creative. *The Chronicle of Higher Education, 53,* B11.

Sternberg, R. J., & The Rainbow Project Collaborators. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence, 34,* 321–350.

Received January 23, 2008
Revision received March 2, 2009
Accepted March 11, 2009 ■

Journal of Educational Psychology
2009, Vol. 101, No. 4, 982–997

0022-0663/09/$12.00 DOI: 10.1037/a0016351

Development and Validation of a Measure of Academic Entitlement: Individual Differences in Students' Externalized Responsibility and Entitled Expectations

Karolyn Chowning and Nicole Judice Campbell
University of Oklahoma

Four studies present the validation of a self-report scale capturing *academic entitlement*, which is defined as the tendency to possess an expectation of academic success without a sense of personal responsibility for achieving that success. The Academic Entitlement scale possesses a 2-factor structure (Study 1): 10 items measure students' Externalized Responsibility for their academic success, and 5 items measure students' self-serving Entitled Expectations about professors and course policies. In Study 2, the Externalized Responsibility subscale correlated positively with related measures of entitlement, grandiosity, and narcissism, and it was negatively related to self-esteem, personal control, need for cognition, agreeableness, and conscientiousness. In Study 3, participants rated various responses to academic situations selected by university instructors as highly inappropriate or highly appropriate. The Academic Entitlement scale predicted students' ratings of the appropriateness of these student behaviors as well as the likelihood that they themselves would engage in these behaviors. In a laboratory setting, individuals with high Academic Entitlement scores evaluated the researcher more negatively than those with low Academic Entitlement scores (Study 4). Practical applications are discussed.

Keywords: entitlement, narcissism, student incivility, expectations, scale validation

Student incivility is a phenomenon regularly encountered by university instructors (Amada, 1999; Boice, 1996; Meyers, 2003; Tiberius & Flak, 1999; Tom, 1998). Uncivil student behaviors during lecture include reading a newspaper, talking, answering mobile phones, sending wireless messages, arriving late to class, leaving class early, and inappropriate use of laptop computers in class. Uncivil student behaviors also are evidenced in student–instructor interactions, such as e-mails, calls, and face-to-face conversations that are demanding, too informal, or presumptuous. Identifying the sources of student incivility will allow both researchers and instructors to better understand and address these behaviors in an effective and professional way.

Certain situational factors contribute to collegiate incivility. First-year college students, for example, are likely experiencing a myriad of daily stressors because of academic and social adjustments accompanying the transition from high school to college (Kerr, Johnson, Gans, & Krumrine, 2004; Santiago-Rivera & Bernstein, 1996); these increased levels of stress in turn affect student behavior (e.g., Felsten & Wilcox, 1992). For some students, incivility may be due to a failure to adequately adjust their academic expectations from high school (McGlynn, 2001). Moreover, large lecture sections can foster feelings of perceived anonymity, which may lead to diffusion of responsibility (e.g., Diener, Lusk, DeFour, & Flax, 1980; Wulff, Nyquist, & Abbott, 1987). Students may thus behave especially poorly in large impersonal classes, engaging in distracting, uncivil behavior more frequently than in smaller courses (Carbone, 1998, 1999).

Individual differences—such as attitudes, gender, and personality—also predict student incivility. For example, Gump (2006) reported that students' attitudes regarding the importance of attendance predicted their actual attendance, suggesting that students' attitudes and perceptions play an important role in predicting their behavior. Gender can also play a major role in the classroom (e.g., Constantinople, Cornelius, & Gray, 1988; Crawford & MacLeod, 1990). In one study, the gender of both the student and the professor affected students' timely completion of a course requirement. Specifically, male students delayed completion more often than female students did overall when the professor was female (Louie & Tom, 2005).

Moving beyond attitudes and gender, the purpose of this research is to identify a stable individual difference in personality thought to play a key role in student incivility. Specifically, *academic entitlement*—defined as the tendency to possess an expectation of academic success without taking personal responsibility for achieving that success—is proposed to be a significant contributing factor to students' inappropriate behavior.

Students possessing high levels of the proposed construct academic entitlement are likely to report other personality characteristics that may contribute to their inappropriate academic behavior. The Big Five Personality Inventory (John, Donahue, & Kentle, 1991) captures fundamental aspects of personality: agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience. We predicted that academically uncivil students would report low levels of agreeableness and conscientiousness and, thus,

Karolyn Chowning, Center for Independent and Distance Learning, University of Oklahoma; Nicole Judice Campbell, Department of Psychology, University of Oklahoma.

Correspondence concerning this article should be addressed to Nicole Judice Campbell, University of Oklahoma, Department of Psychology, 455 West Lindsey #705, Norman, OK 73019. E-mail: njudice@ou.edu

that academic entitlement would correlate negatively with these variables. Relationships between academic entitlement and the other three aspects of personality were not expected.

Our working definition of academic entitlement harkens to an externalized locus of control, as students abdicate responsibility for their own academic outcomes. Beliefs about personal control have been widely studied in psychology (see Shell & Husman, 2008, for a recent summary). Weiner's attribution theory of motivation outlines three dimensions of students' causal attributions about their academic successes and failures: control, locus, and stability (Graham & Weiner, 1996). Following these dimensions, the academically entitled students would perceive little control over their performance outcomes if they viewed grades as bestowed upon them by an external source.

Attributions for success and failure may vary on the basis of students' expectations and their outcomes. Participants in several studies were most likely to attribute their performance to stable, internal outcomes (such as ability) when they expected and received a successful outcome; they were most likely to attribute their performance to variable, external outcomes (such as luck) when they did not expect, yet received, an unsuccessful outcome (Feather, 1969; Feather & Simon, 1972; Möller & Köller, 2000). Various attributional strategies are used by individuals motivated to protect their self-worth. In the face of failure, using a self-serving attributional bias by externalizing responsibility for academic performance may help to protect the self from threatening information (e.g., Ross & Nisbett, 1991).

Longitudinal research indicates not only that college students' externality has increased over the second half of the 20th century (Twenge, Zhang, & Im, 2004) but also that internality is a stronger predictor than self-esteem of academic achievement (Crosnoe & Huston, 2007; Flouri, 2006; Stupnisky et al., 2007). Students with an external locus of control also report a greater degree of stress from academic stressors than internalizing students (Abouserie, 1994). Possession of an external locus of control, especially with regard to one's personal experiences, can produce interpersonal derision as well as poor academic outcomes. For example, students' attitudes toward a teacher in a controlled experiment were lower when their manipulated expectation of failure was high (rather than low) and when students reported an external (rather than internal) locus of control (Feldman, Saletsky, Sullivan, & Theiss, 1983). Externalizing participants report higher levels of aggression (Williams & Vantress, 1969), are more indiscriminate when aggressing (Davis & Mettee, 1971), and believe more strongly in others' propensity for aggression (Young, 1992) than participants with an internal locus of control.

This interpersonal profile is reflected in lower levels of the Big Five Personality Inventory's personality factor agreeableness and becomes a major concern in the study of narcissism, an existing individual difference construct similar to academic entitlement. Psychologists typically study "normal" narcissism in nonclinical populations using the Narcissistic Personality Inventory (NPI; Raskin & Hall, 1979, 1981). Narcissism is a multifaceted construct consisting of superiority, self-absorption, authority, and a sense of entitlement. The Entitlement/Exploitiveness (E/E) subscale of the NPI (identified by Emmons, 1987) has been shown to predict aggressive and derogatory interpersonal behaviors. Individuals high on the E/E subscale of the NPI experienced significant boosts to positive affect following downward social comparisons, suggesting that entitlement/exploitiveness in particular may be the aspect of narcissism most relevant for negative interpersonal interactions (Bogart, Benotsch, & Pavlovic, 2004).

Although E/E is the subscale of the NPI most relevant to the current research, psychometric and theoretical concerns have been raised about its validity as a stand-alone measure of entitlement (e.g., Kubarych, Deary, & Austin, 2004). A recent, more valid measure of entitlement is the Psychological Entitlement Scale (PES; Campbell, Bonacci, Shelton, Exline, & Bushman, 2004). High scores on PES predict a host of self-serving behavioral outcomes beyond the E/E subscale of the NPI, including increased perceptions of deserved salary, competitive resource-depleting decisions, self-oriented romantic relationship attitudes, and a lack of forgiveness (Campbell et al., 2004; Exline, Baumeister, Bushman, Campbell, & Finkel, 2004). Among the negative outcomes predicted by existing measures of entitlement, reactions to ego threat are particularly relevant to the current research explaining student incivility. Psychological entitlement specifically predicts aggressive responses to negative academic feedback (Campbell et al., 2004, Study 9).

Why develop a measure of academic entitlement if a psychological entitlement scale has recently been developed? Students who behave in an entitled fashion in their academic coursework may not display this behavior with their peers, family, or health professionals, and they may not internalize more general entitlement statements as applying to them (e.g., "I am entitled to more of everything" from the PES; Campbell et al., 2004). Further, benefits have been found when examining domain specific contexts of broad personality measures, such as self-esteem. Crocker and colleagues have established domains of contingent self-worth that predict different patterns of behavior (Crocker, Luhtanen, Cooper, & Bouvrette, 2003). The Academic Contingencies of Self-Worth subscale, for example, predicts self-evaluation after receiving admissions decisions to graduate school beyond the variance captured by global self-evaluations including self-esteem (Crocker, Sommers, & Luhtanen, 2002). Similarly, the Academic Entitlement (AE) scale can be validated as a useful measure if it predicts outcome variables, such as uncivil student behavior, beyond the PES.

Building on extant entitlement research, in the present research we outline the development of a more specific predictor of inappropriate student behaviors, one that combines elements of entitlement with pertinent areas of the academic domain. Proposing a new construct to explain student incivility requires both overlapping with relevant constructs, such as psychological entitlement, and capturing unique variance pertaining to the academic situation. The construct of academic entitlement captures this combination of individual and situational factors.

In four studies, we develop and validate a self-report scale to measure academic entitlement. First, we narrow potential scale items on the basis of exploratory factor analysis and reliability coefficients, and we assess the relationship of the scale scores with similar published scales to establish convergent validity. In Study 2, we replicate the findings of Study 1, confirming the two-factor structure in another large sample as well as replicating previous correlations. In Study 3, we establish the predictive validity of the measure through its ability to predict participants' responses to academic situations, specifically, the appropriateness of student behaviors and the likelihood they would engage in those behaviors.

In Study 4, we further examine the predictive validity of AE scale scores in a behavioral validation study in which participants respond to an experimenter who has graded their work. Taken together, these studies establish the efficacy and validity of the AE scale.

Study 1: AE Scale Development

Method

Participants

Undergraduate students (N = 453) enrolled in introductory psychology participated in an online self-report study for partial fulfillment of a course requirement. Complete data were obtained for 442 participants (260 women, 182 men), 61.1% of whom were first-year college students.[1] The majority of participants (76%) reported their age as 18 or 19 years of age; an additional 17% were 20–22 years of age, and less than 6% of the sample reported ages ranging from 23 to 46 years of age. Approximately 76% identified their ethnicity as Caucasian, 4% as Black, 3% as Native American, 9% as Asian or Pacific Islander, 4% as Hispanic or Latino/a, and 2% as other.

Materials

Participants completed a series of questionnaires online. These included the NPI (Raskin & Hall, 1979), the PES (Campbell et al., 2004), the State-Trait Grandiosity Scale (Rosenthal, Hooley, & Steshenko, 2003), the Need for Cognition Scale (Cacioppo & Petty, 1982; Cacioppo, Petty, & Kao, 1984), the Rosenberg Self-Esteem Scale (Rosenberg, 1989), a measure of personal control (the Spheres of Control Scale; Paulhus, 1983), and potential AE scale items. Correlations between the newly developed AE subscales and the published scales were examined for convergent and divergent validity.

The *NPI* (Raskin & Hall, 1979) assesses normal narcissism using the 37-item forced choice version with four subscales identified by Emmons (1987). The E/E subscale of the NPI includes entitled choices, such as "I find it easy to manipulate people" and "I insist upon getting the respect that is due me." The *PES* (Campbell et al., 2004) captures the pervasive sense that one deserves more and is entitled to more than others are. The scale consists of nine items scored on a 6-point scale. An example item is "If I were on the Titanic, I would deserve to be in the *first* lifeboat!" The *State-Trait Grandiosity Scale* (Rosenthal et al., 2003) captures narcissistic arrogance and grandiosity without including the classic conceptualization of self-esteem (i.e., Rosenberg, 1989). The scale consists of 16 adjectives—such as "superior," "envied," and "glorious"—scored on a 7-point scale.

The *Spheres of Control Scale* (Paulhus, 1983) captures perceived locus of control in three main spheres of life, one of which is personal achievement. The Personal Control subscale consists of 10 items scored on a 7-point scale. An example item is "I can usually achieve what I want if I work hard for it." The *Need for Cognition Scale* (Cacioppo & Petty, 1982; Cacioppo et al., 1984) captures the tendency to engage in and enjoy effortful cognitive endeavors. The short form consists of 18 items scored on a 6-point scale. An example item is "I really enjoy a task that involves coming up with new solutions to problems." The *Rosenberg Self-Esteem Scale* (Rosenberg, 1989) contains 10 items scored on a 5-point scale that can be used to assess global self-esteem. An example item is "On the whole, I am satisfied with myself." Reliability coefficients for all scales reached acceptable levels (alphas ranged from .66 to .95).

To measure students' sense of academic entitlement, we tested 31 self-report statements. Test items were selected from a larger pool of items generated by our research team and our undergraduate research assistants. Regular meetings, during which the research team shared anecdotes, academic frustrations, and personal classroom experiences, culminated in a group picture of the entitled student. Items were generated in the spirit of capturing opinions and statements of the prototypical entitled student—the muse for this research—who boldly expresses his or her expectations for academic success while clearly absolving him- or herself of responsibility for achieving that success. This working definition consisted of two parts—responsibility and expectations—that characterized the two themes of the items. Example statements include "Most professors don't really know what they are talking about" and "I should never receive a zero on an assignment that I turned in." After compiling a list containing items individually written by each member of the research team, potential items were discussed, edited for precision and clarity, and strongly overlapping or duplicating items were eliminated. Study 1 establishes the factor structure and provides validation for the AE scale.

Results and Discussion

Exploratory Factor Analysis

We conducted an exploratory factor analysis using principal components extraction with quartimax rotation[2] to explore patterns of covariance between potential AE scale items and to assess the degree to which the pool of 31 items measured multidimensional constructs. Principal components analysis was chosen to reduce the 31 potential items to those that represent the underlying components that contribute most to the variance among the items.

Ten factors emerged with eigenvalues larger than one and accounted for 87.15% of the common variability. Two of those factors accounted for almost 40% of the variability, and the scree test indicated that two factors best described the data (Cattell, 1978). None of the other factors accounted for more than 9% of the variance; a two-factor solution was thus considered the most parsimonious explanation of the data.

To retain items that loaded strongly and singularly on their latent factor, we used two methods to eliminate items. First, a factor loading criterion of .4 or higher was used, such that items with standardized coefficients greater than .4 were retained. Of the remaining items, we assessed cross-loadings using a 2:1 ratio, such that items whose standardized coefficient on one factor was more

[1] Study 1 was conducted during the beginning of a spring semester; students identified as first-year students had been enrolled in college for less than 6 months. Study 2, however, was conducted during the beginning of a fall semester; students identified as first-year students had been enrolled in college for less than 1 month.

[2] Although quartimax is an orthogonal rotation, other rotations (including oblique rotations, such as direct oblimin and promax) produce the same items on each subscale.

than twice the standardized coefficient on the other factor were retained. Therefore, items were eliminated for not loading highly on either of the factors or for loading on multiple factors (following Kim & Mueller, 1978). Two factors accounted for 39.20% of the variance in the final 15 items (see Table 1). The first factor accounted for 24.38% of the variance, possessed an eigenvalue of 7.66, and appears to capture students' Externalized Responsibility. The second factor accounted for 14.82% of the variance, possessed an eigenvalue of 4.65, and seems most related to students' Entitled Expectations.

Internal Consistency

To determine whether any of the remaining AE scale items were problematic, we obtained correlations between the response to a particular item and the sum of the responses to all other items for each subscale. Item-total correlations for the Externalized Responsibility subscale ranged from .40 to .58, and from .27 to .51 for the Entitled Expectations subscale. Additionally, Cronbach's coefficient alpha was computed for both subscales to measure reliability. In this sample, $\alpha = .81$ for the 10-item Externalized Responsibility subscale, and $\alpha = .62$ for the five-item Entitled Expectations subscale. The elimination of any one of the items from either of these subscales would not increase the value of Cronbach's alpha for either subscale, so all of the remaining items from the factor analysis were retained.

AE Scale

Fifteen items in two subscales remained from 31 potential AE scale items as a result of exploratory factor analysis loadings and reliability coefficients. The two subscales, Externalized Responsibility and Entitled Expectations, correlated with each other—in this sample, $r(440) = .21$, $p < .001$—but are distinct constructs and thus are not summed together. Means and standard deviations for both subscales in all four studies are provided in Table 2.

The 10-item Externalized Responsibility factor focuses on students' and professors' responsibilities in the learning process and includes items such as "It is ultimately my professors' responsibility to make sure that I learn the material of a course" and "For group assignments, it is acceptable to take a back seat and let others do most of the work if I am busy." The Externalized Responsibility items are worded such that a high score indicates an entitled lack of personal responsibility. Scores on this subscale averaged below the midpoint of the scale and were positively skewed (skewness = 0.79; see Table 2).

The five-item Entitled Expectations factor focuses on students' expectations of professors' policies and grading strategies and includes items such as "Professors must be entertaining to be good" and "My professors should curve my grade if I am close to the next letter grade." A high score on these Entitled Expectations items indicates students' specific, relatively inflexible, entitled expectations about professor behaviors and grades. Scores on this

Table 1
Factor Loadings of the 15-Item Academic Entitlement Scale

Subscale/items	Factor 1: Externalized Responsibility		Factor 2: Entitled Expectations	
	Study 1	Study 2	Study 1	Study 2
Externalized Responsibility subscale				
1. It is unnecessary for me to participate in class when the professor is paid for teaching, not for asking questions.	**.526**	**.639**	.222	.125
2. If I miss class, it is my responsibility to get the notes. *(Reverse)*	**.659**	**.533**	−.099	−.046
3. I am not motivated to put a lot of effort into group work, because another group member will end up doing it.	**.702**	**.679**	−.089	−.074
6. I believe that the university does not provide me with the resources I need to succeed in college.	**.618**	**.672**	.032	.095
7. Most professors do not really know what they are talking about.	**.590**	**.624**	.096	.099
10. If I do poorly in a course and I could not make my professor's office hours, the fault lies with my professor.	**.621**	**.578**	.108	.234
11. I believe that it is my responsibility to seek out the resources to succeed in college. *(Reverse)*	**.516**	**.548**	−.129	.021
12. For group assignments, it is acceptable to take a back seat and let others do most of the work if I am busy.	**.651**	**.719**	.041	.001
13. For group work, I should receive the same grade as the other group members regardless of my level of effort.	**.611**	**.604**	.092	.010
15. Professors are just employees who get money for teaching.	**.599**	**.605**	.129	.119
Entitled Expectations subscale				
4. My professors are obligated to help me prepare for exams.	.101	.126	**.551**	**.551**
5. Professors must be entertaining to be good.	.009	.111	**.495**	**.541**
8. My professors should reconsider my grade if I am close to the grade I want.	.202	.223	**.725**	**.718**
9. I should never receive a zero on an assignment that I turned in.	.039	.025	**.603**	**.700**
14. My professors should curve my grade if I am close to the next letter grade.	.049	.077	**.726**	**.765**

Note. Participants rate each item on a 7-point scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). The first 10 items compose the first subscale, Externalized Responsibility, which captures an entitled lack of responsibility for one's education. The last five items compose the second subscale, Entitled Expectations, which captures students' entitled expectations about professors and their course policies. Boldface indicates that the item loaded on the specified factor more strongly than the other.

Table 2
Descriptive Statistics for the Externalized Responsibility (ER) and Entitled Expectations (EE) Subscales of the Academic Entitlement Scale

	Study 1 (N = 442)		Study 2 (N = 886)		Study 3 (N = 357)		Study 4 (N = 120)	
Variable	ER	EE	ER	EE	ER	EE	ER	EE
Overall	2.26 (0.82)	4.51 (1.02)	2.19 (0.84)	4.41 (1.10)	2.59 (0.88)	4.63 (0.95)	1.94 (0.61)	4.71 (0.99)
Male	**2.52 (0.86)**	4.52 (1.03)	**2.43 (0.89)**	4.41 (1.16)	2.63 (0.83)	4.63 (0.92)	**2.19 (0.59)**	4.57 (1.09)
Female	**2.07 (0.74)**	4.50 (1.02)	**2.04 (0.77)**	4.42 (1.07)	2.55 (0.92)	4.64 (0.98)	**1.85 (0.59)**	4.76 (0.96)
First year	2.29 (0.84)	4.55 (1.02)	**2.15 (0.83)**	4.41 (1.11)	2.61 (0.88)	4.70 (0.98)		
Upper class	2.21 (0.79)	4.46 (1.03)	**2.34 (0.87)**	4.42 (1.08)	2.55 (0.89)	4.52 (0.89)		

Note. Values are means with standard deviations in parentheses. Bolded pairs are significantly different at $p < .05$ (e.g., male and female ER in Study 1). The Academic Entitlement scale items are rated on a 7-point scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

subscale averaged above the midpoint of the scale and were slightly negatively skewed (skewness = −0.21; see Table 2).

Participants rated each item on a 7-point scale from *strongly disagree* to *strongly agree*. AE subscale scores were calculated by reverse-scoring Items 2 and 9 (see Table 1 for items) and then calculating the mean response, which could range from 1 to 7. Scores on the Externalized Responsibility subscale were lower than those on Entitled Expectations subscale (see Table 2 for means).

Demographics

The AE scale measures faulty expectations about college coursework, so one might expect AE scores of first-year college students to be higher than scores of upper class students. However, no differences were found between first-year (see Footnote 1) participants (n = 270) and upper class students (n = 172) on either subscale of the AE scale (see Table 2). This provides some support for the idea that academic entitlement as an individual difference does not necessarily lessen with increased experience in college.

Researchers and clinicians working with narcissism report a gender difference, such that men report higher agreement with the narcissistic choices of the NPI (Raskin & Hall, 1979) and are more frequently diagnosed with narcissistic personality disorder than women (American Psychiatric Association, 2000). Consistent with past findings, in this sample of 260 female and 182 male undergraduates, male students had slightly higher scores than female students on the NPI (M = 14.83, SD = 6.33 vs. M = 13.63, SD = 6.06, respectively), $t(411) = -1.95$, $d = 0.19$, $p = .05$. This gender difference was not present in this sample for the PES (as found in previous research; Campbell et al., 2004) or the Entitled Expectations subscale of the AE scale (both ts < 1).

For the Externalized Responsibility subscale, in which high scores indicate an entitled lack of responsibility, male students had significantly higher scores than female students, AWS $t'(352.35) = -5.83$, $d = 0.56$, $p < .001$ (see Table 2).[3] These findings are consistent with previously identified gender differences in variables related to narcissism and entitlement, as well as student success, particularly in the first year of college. Similar to PES, no gender difference was found for the Entitled Expectations subscale, and similar to NPI, male students reported higher agreement with the Externalized Responsibility subscale.

Construct Validity

The AE subscales correlated with related measures (see Table 3). The Entitled Expectations subscale correlated positively with psychological entitlement and negatively with need for cognition. The Externalized Responsibility subscale correlated negatively with personal control, need for cognition, and self-esteem—constructs construed as adaptive in an academic environment. The negative correlation with personal control in particular is sizeable for self-report measures of personality. By indicating a significant relationship between external locus of control and academic entitlement, this correlation directly helps to confirm the characterization of the subscale as Externalized Responsibility.

Externalized Responsibility also correlated positively with narcissism-related constructs, including psychological entitlement, the E/E subscale of the NPI, and grandiosity. Given the wide differences between the context and content of the AE scale items compared with these four narcissism-related measures, small correlation coefficients were expected (see Table 3). The two AE subscales are clearly distinct from related measures providing evidence of discriminant validity.

The pattern of results for the Externalized Responsibility subscale follows that of psychological entitlement (PES) and narcissistic entitlement/exploitiveness (E/E), with several notable exceptions. First, Externalized Responsibility correlated negatively with personal control (see Table 3), whereas other measures of entitlement did not (for PES and E/E, rs < .07), indicating that an external locus of control is a key component of academic entitlement. That is, the more personal control a student perceives, the less academic entitlement he or she exhibits. Further, the negative correlation with self-esteem suggests that academic entitlement might be a compensatory, protective strategy separate from narcissism, which typically has a small positive correlation with self-esteem—in this sample, $r(408) = .18$, $p < .001$. Additionally, the negative relationship between the Externalized Responsibility

[3] Because of the unequal sample sizes by gender, the AWS t' was used to correct for the possible inequality of population variances when Levene's test for equality of variances was significant. The Aspen–Welch–Satterthwaite modification to the two-independent-sample t-test adjusts the degrees of freedom and does not assume equal population variances (Toothaker & Miller, 1996).

Table 3
Correlations Between the Externalized Responsibility (ER) and Entitled Expectations (EE) Subscales of the Academic Entitlement Scale and Related Variables Across Two Samples

Measure	Study 1 (N = 442)			Study 2 (N = 886)		
	Academic Entitlement scale			Academic Entitlement scale		
	α	ER	EE	α	ER	EE
Narcissistic Personality Inventory (NPI)	.83	.18**	.10	.84	.09*	.18**
NPI: Entitlement/Exploitiveness subscale	.66	.29**	.06	.68	.22**	.18**
Psychological Entitlement Scale	.86	.38**	.18**	.88	.28**	.34**
State-Trait Grandiosity Scale	.95	.26**	.08	.95	.13**	.17**
Personal Control subscale of the Spheres of Control Scale	.73	−.47**	−.03	.73	−.52**	−.07
Rosenberg Self-Esteem Scale	.86	−.24**	−.08	.88	−.28**	−.01
Need for Cognition Scale	.90	−.21**	−.16**	.79	−.25**	−.18**
Big Five Personality Inventory						
Agreeableness				.79	−.35**	−.01
Conscientiousness				.78	−.38**	−.08
Extraversion				.87	−.18**	.03
Openness to experience				.80	−.11	−.02
Neuroticism				.84	.11*	.02

* $p < .01$. ** $p < .001$.

subscale and need for cognition (see Table 3) is reduced—for PES, $r(421) = -.17$, $p < .001$—or eliminated—for E/E, $r < .04$—for other measures of entitlement, suggesting that academic entitlement captures a more cognitive or academic construct than these broader measures, as intended.

In Study 1, a 15-item AE scale was developed, consisting of two subscales measuring an entitled lack of Externalized Responsibility and faulty, self-serving Entitled Expectations. The subscales were internally consistent and related to, yet distinct from, similar constructs. Additionally, the Externalized Responsibility subscale was negatively related to constructs capturing positive attributes, such as personal control, self-esteem, and need for cognition. In Study 2, these findings are replicated in a larger sample.

Study 2: AE Scale Replication

Method

Participants

Undergraduate students (n = 911) enrolled in introductory psychology received credit toward completion of a course requirement for participating in an online self-report study. Complete data were obtained for 886 participants (551 women, 335 men), 72.2% of whom were first-year college students (see Footnote 1). The majority of participants (96%) reported their age as 18–22 years of age, and less than 4% of the sample reported ages ranging from 23 to 55 years of age. Approximately 75% identified their ethnicity as Caucasian, 6% as Black, 5% as Native American, 6% as Asian or Pacific Islander, 3% as Hispanic or Latino/a, and 2% as other.

Materials

Participants completed a series of questionnaires online, including all scales from Study 1, as well as the Big Five Personality Inventory (John et al., 1991) and the new 15-item AE scale. The Big Five Personality Inventory assesses openness, conscientiousness, extraversion, agreeableness, and neuroticism. The scale consists of 44 adjective phrases, such as "is talkative," "can be cold and aloof," and "gets nervous easily"; participants rate on a 5-point scale their agreement with the statement as descriptive of them. Reliability coefficients for all scales reached acceptable levels (see Table 3).

Results and Discussion

Factor Structure

For the 15 items identified in Study 1 as the AE scale, a two-factor solution again fit the data using principal components analysis with quartimax rotation (see Table 1). The first factor, Externalized Responsibility, possessed an eigenvalue of 8.49, which accounted for 26.75% of the variance in the 15 items. The second factor, Entitled Expectations, possessed an eigenvalue of 5.00 and accounted for 15.76% of the variance. Additionally, the Externalized Responsibility (α = .83) and Entitled Expectations (α = .69) subscales remained internally consistent and correlated with each other, $r(884) = .25$, $p < .001$. Interitem correlations within the two factors further support the fit of the two-factor model (i.e., all correlations were significant at $p < .001$), and the reliability of each subscale did not increase with the deletion of any items.

These results are comparable with those obtained in Study 1; the second exploratory factor analysis provided the same factor structure as the first. To provide a more powerful test of the hypothesized factor structure (Gorsuch, 1983), and to further establish that the proposed two-factor structure provides a good fit for the data, we conducted confirmatory factor analyses using the SAS function PROC CALIS. Specifically, the proposed two-factor model was tested against a one-factor alternative. We compared the fit of each model to the data using four fit indices: chi square, the goodness-of-fit index (GFI; Joreskog & Sorbom, 1996), the comparative fit

index (CFI; Bentler, 1990), and the root-mean-square error of approximation (RMSEA; Steiger, 1990).[4] These indices allow a model to be tested not only against a comparative model but also against an objective standard.

Both models met convergence criteria (i.e., there were no convergence issues). As predicted, the two-factor model provided a better fit to the data. The one-factor model tested the assumption that the 15 items all reflect a single underlying latent construct of academic entitlement. The one-factor model provided only a marginally acceptable fit to the data (χ^2 = 1,040.68, GFI = .840, CFI = .696, RMSEA = .109). The two-factor model tested the proposed scale structure, with items loading on previously identified factors of Externalized Responsibility and Entitled Expectations (as specified in Table 1). The two-factor model provided a good fit for the data (χ^2 = 410.08, GFI = .938, CFI = .897, RMSEA = .064). Moreover, all four fit indices that were used yielded a better fit for a two-factor model when compared with the competing one-factor model, thus supporting the two-factor model as the formal measurement model for the AE scale.

Construct Validity

Following the results of Study 1, the AE subscales again correlated with related measures as expected (see Table 3). The Externalized Responsibility subscale was positively correlated with psychological entitlement, the E/E subscale of the NPI, and grandiosity. Externalized Responsibility again was negatively correlated with educationally adaptive constructs, including personal control, need for cognition, and self-esteem. Data for the Big Five Personality Inventory were also collected in this sample. Both agreeableness and conscientiousness were negatively correlated with Externalized Responsibility and unrelated to Entitled Expectations (see Table 3). The other three Big Five Personality Inventory measures (neuroticism, extroversion, and openness to experience) were unrelated to both subscales (rs < |.12|). The Entitled Expectations subscale correlated positively with psychological entitlement as well as narcissism and E/E. Again, Entitled Expectations is not linked with as many published scales as Externalized Responsibility, but it has a unique contribution to explaining inappropriate student behavior, as is demonstrated in Study 3.

Demographics

Replicating Study 1, scores on the Externalized Responsibility subscale were lower than those on Entitled Expectations subscale (see Table 2). Items composing the Externalized Responsibility subscale are worded such that high scores indicate an entitled lack of responsibility, so students' mean response of 2.19 to this subscale is closer to the unentitled *strongly disagree* pole.

Unlike Study 1, a difference emerged between first-year and upper class students in the Externalized Responsibility subscale. First-year participants had significantly lower (less entitled, more responsible) scores than upper class students, AWS $t'(315.39) = -2.44$, $d = 0.22$, $p = .015$ (see Table 2 and Footnote 3). This finding may be due to semester differences; in Study 2, first-year students had just begun their first semester in college, but in Study 1, first-year students were beginning their second semester. This difference suggests that some experience with a college environment may be necessary to exacerbate students' perceived lack of control over their educational outcomes. No difference was found for the Entitled Expectations subscale.

Consistent with Study 1, gender differences were also found for the Externalized Responsibility subscale in Study 2; here again, male students had significantly higher scores than female students, AWS $t'(625.92) = -6.75$, $d = 0.47$, $p < .001$ (see Table 2 and Footnote 3). As expected, no gender differences were present in this sample for Psychological Entitlement or the Entitled Expectations subscale of the AE scale.

In both Study 1 and Study 2, the factor structure, internal consistency, and construct validity of both subscales of the AE scale were established. Specifically, a 10-item subscale capturing Externalized Responsibility and a 5-item subscale capturing Entitled Expectations were established in exploratory factor analyses and confirmed in confirmatory factor analyses. Moreover, the subscales correlated as predicted with related measures, suggesting that the AE subscales capture the constructs intended. To further establish the utility of the AE scale, in Study 3 we provide a test of the predictive validity of subscale scores.

Study 3: Predicting Students' Predicted Behavior

In addition to the AE scale's relationships with published constructs, this newly developed scale can be used to predict students' own ratings of inappropriate and appropriate student behaviors. A vignette measure was created to assess students' perceptions of different academic behaviors. In four vignettes about academic situations, students rated multiple response options identified by instructors as inappropriate or appropriate. The participants rated both the appropriateness of each behavior and the likelihood they themselves would engage in the behavior.

Method

Participants

Undergraduate students (N = 386) enrolled in introductory psychology received credit toward completion of a course requirement for participating in an online self-report study. Complete data were obtained for 357 participants (173 women, 184 men), 63.3% of whom were first-year college students. Approximately 82% identified their ethnicity as Caucasian, 5% as Black, 4% as Native American, 4% as Asian or Pacific Islander, 3% as Hispanic or Latino/a, and 1% as other. Although age information was not collected in this sample, the sample was drawn from the same university population (albeit a different semester) as the first two studies and is thus likely similar.

Procedure

Participants completed a series of questionnaires online, including measures of psychological entitlement and personal control,

[4] For the chi-square statistic, small values that fail to reach significance indicate a good fit. However, chi-square is very sensitive to sample size, rendering it unclear in many situations whether statistical significance is due to poor fit of the model or to the size of the sample. This uncertainty has led to the development of many other statistics to assess overall model fit, including CFI, GFI, and RMSEA. Values for these three measures range from 0 to 1; larger values indicate better fit for both the CFI and GFI. RMSEA measures error in the model, thus smaller values indicate better model fit.

the AE scale, and an additional scale measuring strategic flexibility. The Strategic Flexibility Scale (Cantwell & Moore, 1996) consists of 21 items in three subscales. The Adaptive subscale identifies the use of flexible approaches to college work; the Inflexible subscale captures unwillingness to adjust their academic strategies; and the Irresolute subscale measures a lack of academic strategies. The scales collected possessed reliability coefficients ranging from .66 to .88.

Student Behavior Vignettes Measure

To ensure the participants were operating in similar stimulus space, we developed a vignette measure to identify specific uncivil student behaviors. We generated academic scenarios thought to evoke entitled behaviors and, in a previous sample, collected student responses to open-ended questions about the scenarios. Participant-generated statements that appeared to capture a continuum of student responses were selected and retained to administer to participants in the current study.

The vignette measure consisted of four vignettes describing academic situations including exam preparation, homework policies, beliefs about general education courses, and course grades (see Table 4). For example, "In one of your classes this semester, you check your grade and see that it is just below the cutoff for a higher grade." Each vignette was followed by five to nine responses per situation, such as "If I came to class every time and tried, I think I should get an A" and "I would deserve the grade I earned throughout the semester. I could have tried harder to get a higher grade." Students rated each of these responses with respect to the likelihood they would make this statement or engage in this behavior using a 6-point scale ranging from *highly unlikely* to *highly likely*. Next, students rated all responses regarding the appropriateness of each statement or behavior on a similar 6-point scale ranging from *highly inappropriate* to *highly appropriate*.

To establish an objective standard for the appropriateness of the student behavioral response options, subject-matter experts rated the vignette responses on appropriateness. The experts were instructors ($N = 21$) recruited from the psychology department, with teaching experience ranging from several months to 37 years.

Results and Discussion

Student Behavior Vignettes Scores

The vignette items used in subsequent analyses were selected on the basis of subject-matter expert rater consensus. Items rated by subject-matter experts as highly inappropriate ($M < 2$, 14 items) or highly appropriate ($M > 5$, 11 items) were identified (see Table 5).

An exploratory factor analysis was then conducted on the student responses to these 25 identified items ($N = 357$). Principal components extraction with a quartimax rotation was used to assess the relationships between them. A scree plot of the eigenvalues indicated that two dominant factors, reflecting inappropriate and appropriate student responses, did in fact underlie responses to the 25 vignette items. The inappropriate and appropriate items possessed a reliable factor structure, as indicated by their respective eigenvalues and Cronbach's alphas. The 14 inappropriate responses included items such as "I would complain to the professor who misled me!" (a response in the exam preparation vignette; $\lambda = 4.63$, Cronbach's $\alpha = .86$). The 11 appropriate responses included items such as "I would answer the questions to the best of my ability" (also a response in the exam preparation vignette; $\lambda = 5.35$, Cronbach's $\alpha = .81$).

Participants as a whole were able to distinguish between inappropriate and appropriate responses ($M = 2.43$, $SD = 0.79$ vs. $M = 3.53$, $SD = 0.81$, respectively), $t(383) = -16.07$, $d = 1.37$, $p < .001$. The distinction between inappropriate and appropriate responses was stronger for participants with low scores (<-1 *SD*) on both AE subscales ($M = 1.65$, $SD = 0.68$ vs. $M = 4.17$, $SD = 0.74$, respectively), $t(19) = -12.58$, $d = 3.55$, $p < .001$. Interestingly, the difference in *appropriateness* ratings between inappropriate and appropriate responses failed to reach significance for participants with high scores (>1 *SD*) on both AE subscales ($M = 3.16$, $SD = 1.02$ vs. $M = 3.17$, $SD = 0.75$, respectively), $t(6) = -0.01$, *ns*. This suggests that academic entitlement is related to knowledge of appropriate student behaviors in a university setting.

Each set of items was averaged to form four scores: students' *likelihood* ratings of inappropriate items (i.e., students' estimated likelihood that they themselves would make this statement or engage in this behavior; $M = 2.64$, $SD = 0.69$), students' *likelihood* ratings of appropriate items ($M = 3.27$, $SD = 0.71$), students' *appropriateness* ratings of inappropriate items (i.e., students' ratings of the appropriateness of making this statement or engaging in this behavior; $M = 2.41$, $SD = 0.79$), and students' *appropriateness* ratings of appropriate items ($M = 3.56$, $SD = 0.81$).[5] These four scores were then predicted by related measures, including the AE subscales.

Predicting Appropriate Items

Multiple regression analyses were conducted to predict students' ratings of the *likelihood* of appropriate items and students' ratings of the *appropriateness* of appropriate items. In addition to Externalized Responsibility and Entitled Expectations scores, related variables—such as personal control (Paulhus, 1983), strategic flexibility (Cantwell & Moore, 1996), and psychological entitlement (Campbell et al., 2004)—were considered as predictors. Zero-order correlations for all variables included in Study 3 are presented in Table 5.

When predicting the *likelihood* of appropriate items, two significant variables accounted for half of the variance, $R^2_{adj} = .51$, $F(2, 353) = 186.36$, $p < .001$. Students' ratings of the *likelihood* of appropriate items were strongly related to students' *appropriateness* ratings for appropriate items, $\beta = .633$, $t(355) = 14.82$, $p < .001$, and were negatively related to the Externalized Responsibility subscale of the AE scale, which captures an entitled lack of responsibility, $\beta = -.147$, $t(355) = 3.44$, $p < .001$. Thus, the AE scale retained a significant relationship with students' self-reported *likelihood* of engaging in appropriate academic behaviors—even

[5] As the word appropriate is used to label both a type of prompt and a type of response, italics and underlines are used throughout. *Likelihood* and *Appropriateness* indicate students' ratings of their perceived likelihood of engaging in the response and their ratings of the appropriateness of the responses (see Table 4). Inappropriate and Appropriate, however, indicate subject-matter experts' consensus ratings of the appropriateness of the responses.

Table 4
Student Behavior Vignettes With Appropriate and Inappropriate Responses

Vignette	Responses
In one of your classes this semester, you check your grade and see that it is just below the cutoff for a higher grade.	1. I would not expect a grade change. My idea of a college professor is not that lenient. 2. *I think the higher grade would be fair. I am very concerned about my GPA, and a lot of my teachers in high school were somewhat flexible with grades.* 3. *I would expect the professor to be a kind, gentle, understanding person and bump me up.* 4. I would deserve the grade I earned throughout the semester. I could have tried harder to get a higher grade. 5. I expect the professor to be fair and honest in his or her grading and make no exceptions for anyone. 6. *If I came to class every time and tried, I think I should get an A.*
You are planning to enroll in a 1000-level introductory course in a subject you are not very interested in. The class is, however, a general education requirement and you need to take the course for your chosen major.	7. *It should be a general class, nothing hard or requiring a lot of work. A class I do not want to take making me work hard would make me never want to take a class in that subject again.* 8. *I expect that the class covers uninteresting, needless information that I will probably never use, and the professor and the class feel the same way.* 9. *I expect that there will be study sessions, very little homework, and a noncomprehensive final.* 10. I expect the professor to teach the class just like he or she would teach any other class. 11. *I expect the professor to make class easy with a good chance of getting an A, and to stick to the book so students who miss class can keep up.* 12. *There should be extra credit opportunities. The class should not hurt your GPA because it is only general education.*
In a humanities course, the professor regularly discusses material not in the assigned readings. As you are taking the midterm, you realize that several of the questions are not from the textbook (which you studied). Although it was likely covered in class, you studied only the assigned readings and thus will not perform as well as you had expected on this exam.	13. *I would be bitter about it. If I did study all of the readings that were assigned I should be given the opportunity to do relatively well on the exam.* 14. There is nothing wrong with a professor having material from class on the exam. I should have taken better notes! 15. I would answer the questions to the best of my ability. 16. *I would be angry—the students are being penalized for missing class even if they had to miss.* 17. I would feel stupid. Always take notes when the professor lectures, it will most likely be on the test. 18. *That is stupid. The "teacher" should be upfront on whether he/she will teach from the book or not and he/she should tell you what will be on the tests.* 19. *It is difficult to try to decide what is trivial "for-your-information" material in most lectures. The testable material should come from a reliable textbook source only.* 20. *I would complain to the professor who misled me!* 21. A bad grade would likely change my view as to how I prepared for the next exam.
Your professor has specific policies about how homework is to be completed. These policies include type of font, font size, margins, stapled pages, and color of ink. You receive a zero on your second homework because you did not staple your homework.	22. I would use the correct format from then on. The policy is ridiculous, but the teacher is free to grade however he or she likes. Besides, it is not very hard to comply with. 23. I would be annoyed at myself. If directions are explained, even if they are picky, you have to go along with them. 24. *I would be mad that something so menial and pointless found at some random place in the syllabus would give me a zero.* 25. Although I obviously would not like it, I would still think it was fair. It is my fault that I messed up; I knew what to do and did it wrong.

Note. Inappropriate responses are italicized. Participants rate all 25 responses on a 6-point scale ranging from 1 (*highly unlikely*) to 6 (*highly likely*) in response to the following prompt: "In this situation, how *likely* is it that you would make the following statements/thoughts?" Next, participants rate all 25 responses on a 6-point scale ranging from 1 (*highly inappropriate*) to 6 (*highly appropriate*) in response to the following prompt: "In this situation, how *appropriate* is it for you to make the following statements/thoughts?" This produces four scores: students' ratings of the *likelihood* of inappropriate items, students' ratings of the *likelihood* of appropriate items, students' ratings of the *appropriateness* of inappropriate items, and students' ratings of the *appropriateness* of appropriate items. GPA = grade point average.

after accounting for students' own ratings of the *appropriateness* of the behaviors.

In a separate analysis predicting students' ratings of *appropriateness* for appropriate items, four significant variables accounted for one third of the variance, $R^2_{adj} = .32$, $F(4, 351) = 42.50$, $p < .001$. Students' *appropriateness* ratings for appropriate items were negatively related to the Externalized Responsibility subscale of the AE scale, $\beta = -.382$, $t(355) = 7.09$, $p < .001$, and were positively related to inflexible academic strategies, $\beta = .248$, $t(355) = 5.16$, $p < .001$. Further, personal control, $\beta = .188$, $t(355) = 3.47$, $p < .01$, and adaptive academic strategies, $\beta = .136$, $t(374) = 2.80$, $p < .01$, emerged as significant predictors of students' *appropriateness* ratings of appropriate items. In addition, then, to significantly predicting students' self-reported *likelihood*

Table 5
Correlations Between Vignette Measures and Constructs Used as Predictor Variables in Study 3

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Likelihood of appropriate items	*.78*										
2. Appropriateness of appropriate items	**.70**	*.86*									
3. Likelihood of inappropriate items	**−.33**	−.23	*.82*								
4. Appropriateness of inappropriate items	**−.27**	**−.42**	**.69**	*.87*							
5. Externalized Responsibility subscale	**−.46**	**−.49**	**.29**	**.32**	*.87*						
6. Entitled Expectations subscale	−.17	−.05	**.56**	**.43**	.16	*.72*					
7. Psychological Entitlement Scale	−.17	−.20	**.30**	**.33**	**.30**	**.25**	*.88*				
8. Personal Control subscale of the Spheres of Control Scale	**.39**	**.43**	−.23	−.23	**−.58**	−.12	−.19	*.78*			
9. Adaptive subscale of the Strategic Flexibility Scale	.08	.05	−.06	.01	.02	−.13	.10	.09	*.68*		
10. Inflexible subscale of the Strategic Flexibility Scale	.17	.21	.20	.16	−.02	**.26**	.00	.03	**−.41**	*.79*	
11. Irresolute subscale of the Strategic Flexibility Scale	−.07	−.10	**.37**	**.25**	.22	**.28**	.04	**−.32**	−.16	**.35**	*.76*

Note. All *r*s greater than .13 are significant at $p < .01$. All *r*s greater than .17 are significant at $p < .001$. Correlations greater than .25 are bolded for visibility. Reliability coefficients (alphas) are italicized and reported on the diagonal.

of engaging in appropriate academic behaviors, entitled responses on the Externalized Responsibility subscale were also the strongest predictor of students' ratings of *appropriateness*, beyond students' adaptively flexible academic strategies and an internal locus of control. Thus, the AE scale predicted students' ratings of appropriate, civil student behaviors. An even stronger test of the scale involves students' ratings of inappropriate, uncivil student behaviors.

Predicting Inappropriate Items

The same procedure of multiple regression analyses was used to predict students' ratings of the *likelihood* of inappropriate items and students' ratings of the *appropriateness* of inappropriate items. The Externalized Responsibility and Entitled Expectations subscales of the AE scale, personal control (Paulhus, 1983), strategic flexibility (Cantwell & Moore, 1996), and psychological entitlement (Campbell et al., 2004) were considered as predictors.

When predicting the *likelihood* of inappropriate items, three significant variables accounted for over half of the variance, $R^2_{adj} = .58$, $F(3, 353) = 162.59$, $p < .001$. Students' ratings of the *likelihood* of inappropriate items were positively related to students' *appropriateness* ratings for inappropriate items, $\beta = .523$, $t(355) = 13.54$, $p < .001$; the Entitled Expectations subscale of the AE scale, $\beta = .291$, $t(355) = 7.47$, $p < .001$; and irresolute academic strategies, $\beta = .157$, $t(355) = 4.32$, $p < .001$. Thus, for inappropriate behavior, the Entitled Expectations subscale of the AE scale accounted for students' self-reported *likelihood* of engaging in inappropriate academic behaviors—even beyond students' own ratings of the *appropriateness* of these inappropriate behaviors. This suggests that together, students' misperceptions of the appropriateness of inappropriate behavior, their maladaptive Entitled Expectations about professors and course policies, and their confusion about which academic strategies to use predict student incivility.

In a separate analysis predicting students' ratings of *appropriateness* for inappropriate items, the AE subscales accounted for one quarter of the variance, $R^2_{adj} = .25$, $F(2, 353) = 58.02$, $p < .001$; they remained the only significant predictors even when all other variables were included in the model. Students' *appropriateness* ratings for inappropriate items were positively related to the Entitled Expectations subscale of the AE scale, $\beta = .387$, $t(355) = 8.27$, $p < .001$, and the Externalized Responsibility subscale of the AE scale, $\beta = .258$, $t(355) = 5.51$, $p < .001$. Thus, both subscales of the AE scale play a role in explaining students' judgments of the appropriateness of inappropriate student behaviors. As addressed earlier, students with entitled scores on both the Externalized Responsibility and the Entitled Expectations subscales judged both inappropriate and appropriate academic behaviors as equally appropriate, indicating entitled students' lack of knowledge about what it takes to succeed in college.

To summarize, in Study 3, the subscales of the AE scale were used to predict students' *likelihood* and *appropriateness* ratings of both inappropriate and appropriate student reactions to academic situations. The AE scale remained a significant predictor in each model, whereas Psychological Entitlement failed to reach significance in any of the four regression models. The strongest predictor of students' *likelihood* ratings was their ratings of *appropriateness*, as expected. The Externalized Responsibility subscale of the AE scale, which indicates an entitled lack of responsibility for one's education, significantly predicted *likelihood* ratings for appropriate items beyond the variance accounted for by *appropriateness* ratings. That is, the strongest predictor of students' predicted behaviors was their ratings of the appropriateness of those behaviors (regardless of their entitlement; regardless also of the valence of those behaviors—both appropriate and inappropriate). Students were more likely to endorse engaging in behaviors that they rated as appropriate.

Beyond the variance accounted for by appropriateness ratings, students' entitled lack of responsibility was the only significant predictor of their estimated likelihood of engaging in appropriate behavior. Students' *appropriateness* ratings were the strongest predictor of student *likelihood* ratings of appropriate behaviors. This is to be expected given the similarity in question structure between vignette measures. That academic entitlement still predicted *likelihood* ratings of appropriate behaviors even beyond students' *appropriateness* ratings provides strong support for the validation of the scale. Students with an externalized sense of responsibility not only rated appropriate items as less appropriate than their nonentitled peers but they also reported being less likely to engage in appropriate behaviors.

Similarly, the Entitled Expectations subscale, which indicates entitled expectations about professors and course policies, significantly predicted *likelihood* ratings for inappropriate items beyond the variance accounted for by *appropriateness* ratings. Students' *appropriateness* ratings were the strongest predictor of student *likelihood* ratings of inappropriate behaviors. This is to be expected given the similarity in question structure between vignette measures. That academic entitlement still predicted *likelihood* ratings of inappropriate behaviors even beyond students' *appropriateness* ratings provides strong support for the validation of the scale. Further, both AE subscales were significant predictors of students' *appropriateness* ratings for both appropriate and inappropriate items. Taken together, these results show the utility of the AE scale in predicting students' predicted reactions to academic situations beyond similar constructs, even the closely related psychological entitlement and personal control.

Thus, Studies 1 and 2 established the factor structure of the scale and the construct validity of the subscales. Study 3 builds on the first two by demonstrating the predictive validity of subscale scores: Scores from the Externalized Responsibility and Entitled Expectations subscales predicted students' reports of their attitudes and predicted behaviors in academic situations presented in vignettes. Taken together, the three studies suggest that the AE scale captures a construct of interest and utility. A limitation of the first three studies, however, is that they assessed only self-reported attitudes and hypothetical behaviors. To address this limitation, in Study 4 we provide behavioral validation by examining relationships between the newly developed scale and students' responses in an experimentally manipulated situation.

Study 4: Behavioral Validation of AE Scale

Reactions to negative feedback should be different for students who possess high levels of academic entitlement compared with nonentitled students. Moreover, the AE scale should predict uncivil student behaviors beyond related constructs, such as psychological entitlement and personal control.

In many ego-threat paradigms, participants are given the opportunity to derogate or aggress against a confederate. In real-world student incivility, however, students derogate the course material and the instructors through their inappropriate behaviors. The AE scale will be best validated in an achievement context as ecologically relevant as possible. Thus, the main dependent variables in Study 4 are participants' evaluations of an academic task they completed and their evaluations of the experimenter who administered and ostensibly scored the academic task.

Method

Participants

Participants were undergraduates (N = 123) enrolled in introductory psychology at a large midwestern state university in the United States. Data were collected across two semesters: early in a fall semester and late in a spring semester. All students received credit toward completion of a course requirement. Complete data were obtained for 120 participants (89 women, 31 men). Further demographics were not collected on this sample; however, participants were drawn from the same student population as the previous three studies and were thus likely similar (i.e., over three quarters were Caucasian; at least two thirds were first-year students, the majority of which were 18–21 years of age).

Procedure

While signing up for the study using an online experiment management system, participants were informed that they would participate in two brief and unrelated studies that were conducted together because of their short length. For the first study, "Personality Correlates," participants were instructed to complete a series of brief individual difference measures online. Upon arriving at the laboratory for the second study, "Academic Task Calibration," participants were told that the researchers were collecting normative data on a task that is highly indicative of academic success and uses verbal ability to measure intelligence, much the same as the ACT. After explanation of the purpose of the task, participants were given a folder that contained the test materials. Prior to their arrival, participants were randomly assigned to receive either no feedback or negative feedback on the academic task.

Participants in both conditions completed two sections of the reading comprehension portion of the Scholastic Aptitude Test. The answer portion of this standardized test was modified from the standard multiple choice to a short-answer essay format. The rationale for this change is two-fold: First, it allowed participants to perceive the grading of this task as more subjective. Second, it allowed the experimenter to include vague comments (in the negative feedback condition) supporting the score on the answer sheet itself. Following the 12-min time limit, the experimenter collected the participants' folders with their responses. Participants in both conditions were told to wait while the experimenter graded their results and entered them into a database. The experimenter left the room for 5 min in both conditions.

Negative feedback condition. After completing the objective academic task, participants in the negative feedback condition received a folder containing a fictitious data sheet explaining the national norms on this test. This sheet also informed participants that they scored in the 33rd percentile on the basis of their performance. In addition to the computer printout of their scores, the folder contained their answer sheet containing gratuitous underlining and circling of their responses, as well as comments from the grader throughout. Specifically, the following phrases were written on the answer sheet itself in red pen: "unclear," "good start," "eh," "close but not quite," and "more . . .".

No feedback condition. After completing the objective academic task, participants in the no feedback condition were told that their data would be analyzed and available at the end of the semester.[6]

Evaluation measures. As the experimenter returned (in both conditions) from ostensibly grading and entering participants' test scores, she explained that the experiment had been selected for

[6] We decided to use a no feedback condition, rather than a positive feedback condition, to avoid interpretation problems. Although use of positive feedback would increase the likelihood of evoking easily measurable effects, it is the impact of negative feedback that we are most interested in when considering entitlement and student incivility. A neutral comparison point, thus, was chosen to avoid interpretation difficulties in discerning the source of condition differences in incivility.

evaluation by the Psychology Department. Specifically, participants were told that the Psychology Department's evaluations of experiments and experimenters are "just like the College of Arts and Sciences' evaluations of courses and professors."

Participants in all conditions completed an evaluation packet consisting of the Positive Affect/Negative Affect Schedule (Watson & Clark, 1994) and an experiment evaluation sheet, on which the participants evaluated both the task and the experimenter. Following Stucke and Sporer (2002), four items addressed the task's validity, accuracy, fairness, and suitability for predicting academic success, and four items addressed the experimenter's usefulness, helpfulness, competence, and accuracy in grading. These eight items were scored on a 7-point scale.

The experimenter began the debriefing process at the close of the experiment. Specifically, participants were asked their thoughts about the purpose of the task and the experiment as a whole. They were further asked about prior exposure to, or knowledge of, the experiment. As a manipulation check, participants were asked how they did on the academic task. The true purpose of the experiment was thoroughly explained, and questions and comments were addressed before participants were thanked and released.

Self-report measures. Participants completed six self-report measures online before arriving for the experiment: AE scale, PES, State-Trait Grandiosity Scale, Personal Control subscale, Contingencies of Self-Worth Scale, and Need for Cognition Scale. Collection of these measures allowed their use in predicting the outcome measures described above.

The Contingencies of Self-Worth Scale (Crocker et al., 2003) captures self-esteem in specific domains in which college students invest their self-esteem; three of these domains are academics, others' approval, and outperforming others. Each subscale consists of five items scored on a 7-point scale. The Academic Competence subscale includes items such as "I feel better about myself when I know I'm doing well academically." The Approval From Others subscale contains items such as "My self-esteem depends on the opinions others hold of me." The Competition subscale is comprised of items such as "I feel worthwhile when I perform better than others on a task or skill." The remaining measures collected in this study are discussed in detail in Study 1.

Results and Discussion

Summary scores were computed for each of the established scales, including academic entitlement. Scores on the Externalized Responsibility subscale and the Entitled Expectations subscale possessed internal consistency ($\alpha = .71$ and $\alpha = .66$, respectively) and were not significantly correlated, $r(118) = .15$, $p = .109$ (see Table 2).[7] Results of reliability analyses for each summary score are reported in Table 6.

Participants' responses on the academic task consisted of 10 short-answer questions. These scores were graded by two trained coders using a 4-point key yielding a score ranging from 0 to 3 per question, in which 0 indicated no answer and 3 indicated the correct answer. Scores between raters were highly correlated, suggesting strong interrater reliability, $r(179) = .981$, $p < .001$. Scores for each question were averaged between raters and totaled to form an overall score for the academic task ($M = 16.85$, $SD = 4.33$). Participants' overall scores for the academic task ranged from 6.5 to 27. Additionally, the number of words participants wrote for each question ($M = 10.21$, $SD = 3.09$) was collected as an estimate of effort and verbal ability. Both overall score and word count were included as potential covariates in regression analyses but failed to reach significance in any of the equations.

Table 6
Correlations Between the Externalized Responsibility (ER) and Entitled Expectations (EE) Subscales of the Academic Entitlement Scale and Related Variables in Study 4

Variable	α	Academic Entitlement scale	
		ER	EE
Self-report variables collected online			
ER	.77		.15
EE	.68	.15	
Psychological Entitlement Scale	.85	.23[b]	.20[b]
Personal Control subscale	.73	−.48[c]	−.17
Academic CSW	.72	−.40[c]	−.01
Others' Approval CSW	.75	.09	.19[b]
Competition CSW	.82	−.09	.07
State-Trait Grandiosity Scale	.94	.14	.13
Need for Cognition Scale	.91	−.14	−.20[b]
Study variables collected in person			
Total score earned on task	.58[a]	−.17	.06
Positive affect (PANAS)	.87	.04	.01
Negative affect (PANAS)	.83	.20[b]	.16
Experiment evaluation	.93	.09	−.07
Experimenter evaluation	.92	−.27[c]	.04

Note. CSW = Contingencies of Self-Worth subscale; PANAS = Positive Affect/Negative Affect Schedule.
[a] The time limit of the task was short so many participants did not complete all of the items; scores for the first five items ($M = 9.13$, $SD = 2.11$) were higher than scores for the second five items ($M = 7.72$, $SD = 3.53$), $t(180) = 4.84$, $p < .001$. [b] Significant at $p = .05$. [c] Significant at $p = .01$.

Scores on Externalized Responsibility (but not Entitled Expectations) differed by gender. Specifically, male students had significantly higher scores than female students on the Entitled Expectations subscale of the AE scale, $t(118) = 2.79$, $d = 0.58$, $p = .006$ (see Table 2). Gender was included in all regression analyses as a potential predictor.

Pearson correlation coefficients were computed between the two subscales of the AE scale and all individual difference measures collected prior to manipulation. As expected, correlations were in the same patterns as previous studies and are presented in Table 6.

[7] This lack of correlation differed by semester. AE scores were related for participants from Spring 2007, $r(33) = .331$, $p = .052$, but unrelated for participants from Fall 2007, $r(83) = .059$, $p = .593$. The AE scale was collected as part of a larger data set in both semesters, so the relationship between Externalized Responsibility and Entitled Expectations was examined in these larger data sets to guard against errors because of small sample size. In prescreening Spring 2007, Responsibility ($M = 2.26$, $SD = 0.82$) and Expectations ($M = 4.51$, $SD = 1.02$) were positively correlated, $r(440) = .207$, $p < .001$. In prescreening Fall 2007, the relationship between Responsibility ($M = 1.94$, $SD = 0.70$) and Expectations ($M = 4.67$, $SD = 1.13$) was still positively correlated, $r(937) = .303$, $p < .001$.

Multiple Regression Analyses

We examined the main hypotheses using a series of multiple regression analyses predicting all dependent variables in the study. Each analysis began with a saturated model that included all available predictor variables. This model was then systematically reduced by eliminating the predictor with the largest p value for its beta weight until all beta weights remaining were significant. Time-order was used to determine the available predictor variables for each outcome variable.

Affect. The Positive Affect/Negative Affect Schedule served as a manipulation check; groups were expected to differ in their negative affect on the basis of the type of feedback (negative or no) that the participants received. Participants receiving negative feedback ($M = 1.57$, $SD = 0.58$) reported significantly higher negative affect than participants receiving no feedback ($M = 1.34$, $SD = 0.38$), $t(162.2) = 3.22$, $d = 0.47$, $p = .002$.

Evaluation. The main dependent variables in the study were participants' responses to the eight evaluation items. The four items in each set were averaged to form scores for evaluation of the experiment (i.e., the task; $M = 3.07$, $SD = 1.38$, $\alpha = .92$) and evaluation of the experimenter (i.e., the person; $M = 6.15$, $SD = 1.08$, $\alpha = .93$). The regression analyses for both evaluation variables included as potential predictors all individual difference measures, gender, experimenter, feedback condition, score earned on the academic task, and affect scores.

The AE scale should be related to more negative evaluations of the experimenter; the Externalized Responsibility subscale or the Entitled Expectations subscale was expected to be the primary predictor, such that more entitled students would evaluate the experimenter more negatively than less entitled students. For the dependent variable evaluation of the experiment, feedback condition (negative, no) should be the only significant predictor, such that students who received negative feedback would evaluate the experiment more negatively than students who did not receive feedback.

As expected, feedback condition, $\beta = -.253$, $t(371) = 3.53$, $p = .001$, was a significant predictor of experiment evaluation, such that participants in the negative feedback condition ($M = 2.73$, $SD = 1.29$) evaluated the task lower than participants receiving no feedback ($M = 3.43$, $SD = 1.39$) $R^2_{adj} = .06$, $F(1, 182) = 12.49$, $p = .001$. No other variable was a significant predictor of participants' evaluation of the experiment (i.e., the task).

The Externalized Responsibility subscale of the AE scale was a significant predictor of experimenter evaluation, such that participants higher in an entitled lack of responsibility evaluated the experimenter (i.e., the person) lower than participants lower in this subscale, $\beta = -.27$, $t(114) = 3.05$, $p = .003$; $R^2_{adj} = .07$, $F(1, 115) = 9.32$, $p = .003$. Again, no other variable was a significant predictor of participants' evaluation of the experimenter (i.e., the person). This relationship was not reduced when potentially confounding variables (gender of participant, experimenter identity; neither significant) were included, $\beta = -.27$, $t(114) = 2.89$, $p = .005$; $R^2_{adj} = .07$, $F(3, 113) = 3.81$, $p = .01$. However, to ensure that feedback condition did not play a role in experimenter evaluation, additional predictors were again considered.

First, condition (recoded as -1 and $+1$ for purposes of the interaction) and the interaction of condition with Externalized Responsibility were added to the reduced model described above. In a model in which feedback condition, Externalized Responsibility, and the interaction between condition and Externalized Responsibility were examined as predictor variables of the criterion variable evaluation of the experimenter, neither feedback condition, $\beta = .01$, $t(114) = 0.13$, $p = .89$, nor the interaction, $\beta = -.02$, $t(114) = 0.21$, $p = .83$, were significant, and the inclusion of feedback condition did not reduce the effect of Externalized Responsibility on experimenter evaluation, $\beta = -.27$, $t(114) = 2.97$, $p = .004$; $R^2_{adj} = .06$, $F(3, 113) = 3.08$, $p = .03$.

The behavioral study provides further validation of the AE scale. After receiving negative academic feedback, participants had higher levels of negative affect and evaluated the academic task lower than participants who did not receive any feedback. Participants with high scores on the AE subscale, representing an entitled lack of responsibility for one's education, evaluated the experimenter (the person, as opposed to the task) lower than participants with low, unentitled scores on Externalized Responsibility.

Additionally, a significant interaction between Externalized Responsibility and feedback condition was not found for test evaluation—that is, students who received negative feedback evaluated the test more poorly than those receiving no feedback, regardless of their AE scale scores. This suggests that the role of AE pertains more to ambiguous or missing feedback—implying that entitled students who score poorly do not complain more loudly than nonentitled students earning low scores. However, regardless of the type of feedback that students received in Study 4, students scoring high on the Externalized Responsibility subscale of the AE scale evaluated the grader more poorly than those scoring low in Externalized Responsibility.

Although a myriad of potential covariates were considered—including other individual differences (academic entitlement, psychological entitlement, grandiosity, need for cognition, personal control, contingencies of self-worth) and task-related characteristics (feedback condition, experimenter, semester, and negative affect)—only the predicted effects of feedback condition and academic entitlement were found for the main outcome variables of experiment and experimenter evaluation.

In sum, Study 4 further establishes the predictive validity of the AE scale. In a laboratory setting, where participants received negative feedback on their test performance, those who scored higher on the AE scale evaluated the experimenter more negatively. This effect mirrors an important manifestation of academic entitlement: student incivility and aggression against evaluators. Finding the effect in a laboratory setting, where one is relatively detached from the evaluation task (compared with a class in which one is enrolled and invested) provides a strong test of the effect. These findings build on and extend the results of the previous three studies; taken together, they present compelling evidence for the validity and utility of the AE scale.

General Discussion

The individual difference of academic entitlement is defined as the expectation of academic success without personal responsibility for achieving that success. A 15-item self-report scale capturing two dimensions of academic entitlement was developed and validated in the current research. The newly developed AE scale

possesses a reliable two-factor structure, with subscales measuring students' Externalized Responsibility for their education and students' Entitled Expectations about professors and their policies.

In Study 1 and Study 2, analyses show that the AE scale possesses a distinct, reliable, two-factor structure and correlates as expected with related constructs. The Entitled Expectations and Externalized Responsibility subscales correlate with measures that capture related constructs, such as psychological entitlement and confusion about academic strategies. The Externalized Responsibility subscale also correlates negatively with scales that capture academically adaptive constructs, such as self-esteem, agreeableness, conscientiousness, need for cognition, and personal control.

In Study 3, the AE subscale scores predicted students' judgments of the appropriateness of various inappropriate and appropriate student behaviors. The appropriateness judgments also predicted students' ratings of the likelihood that they themselves would engage in these inappropriate and appropriate behaviors. Further, the AE scale scores predicted students' self-reported reactions to academic situations designated as inappropriate and appropriate by instructors.

Study 4 demonstrated the predictive ability of the AE scale in a laboratory setting. Participants completed an academic task that consisted of short essay questions; participants who received negative feedback about their performance reported higher levels of negative affect, and they evaluated the academic task lower than participants who did not receive any feedback. Students with high scores indicating an Externalized Responsibility evaluated the experimenter lower than participants with low, unentitled scores on Externalized Responsibility.

Future Directions

These studies provide evidence for the validation of the AE scale. Additional behavioral studies, both in the laboratory and in the classroom setting, could profitably be conducted to further validate the scale. Building on narcissism and ego-threat research (e.g., Smalley & Stake, 1996), differences based on the AE scale can be identified in students' reactions to negative academic feedback, such as a poor grade on a test. Other characteristics of both the student and the course may also play a factor in these reactions. Students' previous academic success, measured both by classroom performance (e.g., grade point average, scores on a formative assessment of the subject matter) and standardized testing (e.g., ACT, GRE, placement examinations), may directly—or indirectly, through academic entitlement—contribute to a portion of the variance in student reactions. Situational characteristics, including the student's major and the applicability of the course or material to their program of study, would likely moderate students' reactions on the basis of their level of academic entitlement. The relationships between the AE scale, students' perceptions of their coursework, and course evaluations should be explored.

In Study 4, academic entitlement was the only significant predictor of participants' ratings of the experimenter, a strong validation of the AE scale. Although these ratings were obtained in a controlled laboratory setting, the experimenters performed some duties comparable with those of an instructor: proctoring and grading an academic task. Therefore, participants' ratings have parallels to end-of-course evaluations of college instructors. To further elucidate the application of this study to the college setting, researchers should examine effects of high levels of academic entitlement on students' expected exam grades and term evaluations of instructors and courses in future research. The findings of this study suggest that students' evaluations of the course will be predicted best by their performance and course grades, but their evaluations of the instructor will be predicted best by their level of academic entitlement.

Instructors' course evaluations are important not only because they play a role in hiring and promotion decisions but because of their role as one of the only evaluative measures of the students' collegiate experience apart from course grades. If these evaluations can be predicted well by individual differences in students' expectations of academic success and students' lack of responsibility for achieving academic success, then the validity of these measures may be called into question. Research on the correlates of student evaluations of instruction may improve the validity of these evaluations by identifying related and confounding measures. This research has the potential to raise instructors' awareness of entitled students' propensity to use a course evaluation as a method of interpersonal aggression. In Study 4, academically entitled students aggressed against the experimenter by evaluating her more negatively than nonentitled students whether they received negative feedback or no feedback about their performance on an academic task. Future research should also examine the behavior of entitled students who receive positive feedback: Does academic success (i.e., positive feedback) suppress the influence of academic entitlement on instructor or experimenter evaluation?

Beyond course evaluations, academic entitlement may have important implications for student retention, success, and graduation. Students who attribute their performance to their courses or instructors may fail to self-correct or develop adaptive strategies to succeed in college. We conceptualize academic entitlement as an individual difference that potentially may be influenced and modified through education and experience. As the majority of participants in these four studies were first-year students, further research is needed to determine whether students' Externalized Responsibility or Entitled Expectations about college-level work changes as they progress through their undergraduate education. Ancedotal evidence from instructors across the curriculum strongly suggests this is not merely a phenomenon among first-year students. However, the relationship of year in school to academic entitlement is unclear, as discussed previously. Thus, additional research using the AE scale is needed to examine longitudinal and cohort effects.

Students who are on academic probation, for example, should have higher levels of entitled attitudes than comparable students in large sections of sophomore-level psychology courses. Students reporting high levels of academic entitlement are using inappropriate academic strategies, do not adjust their expectations about college-level work, and have an external locus of control regarding their academic performance. These attributes are liable to lead to negative outcomes, such as a poor academic record or dissatisfaction with the university. These results may lead to students not returning for a 2nd or subsequent year; academic entitlement may thus be able to explain some of the retention and attrition issues plaguing higher education.

Applications

Although academic entitlement may play a role in student retention outcomes, student incivility is arguably a more salient and pervasive daily concern for instructors. Although the Externalized Responsibility subscale is related to an external locus of control in an academic setting, the Entitled Expectations subscale represents students' faulty expectations about college-level work. Incivility in the classroom is likely due to several factors, including students' failure to adjust their expectations from high school about instructors and classroom policies. Luckily, these faulty expectations may change; individuals no doubt adapt their expectations as a result of experience.

Understanding academic entitlement and its elements can better prepare instructors for dealing with their students. Specifically, preempting tendencies to externalize responsibility and hold unreasonable expectations, instructors can emphasize the student's role in his/her own grade and success, clearly articulating what is expected of the student and what can be expected of the instructor. The quality of instruction can decrease variability in performance on the basis of students' perceptions of control (Perry & Magnusson, 1989).

Moreover, university programs designed to orient students to professors' common standards and practices may remediate maladaptive student expectations. Orientation programs aimed at study skills and time management have been shown to predict improved course performance and increase student retention (e.g., Boudreau & Kromrey, 1994; Sanchez, Bauer, & Paronto, 2006; Ting, Grant, & Plenert, 2000). Attributional retraining in particular has improved the test scores of students with an externalized locus of control (Perry & Penner, 1990). Indices of students' counterproductive expectations, such as academic entitlement, could help to identify incoming students who possess potentially problematic academic strategies and attitudes. Implementation of proactive administrative strategies, such as intrusive advising, may further assist struggling students to obtain the help they need (Earl, 1987; Glennen, Baxley, & Farren, 1985; Schee, 2007).

Student incivility is a problem, especially in larger freshmen-level courses, such as Introductory Psychology. Elucidating sources of student incivility, such as individual differences in academic entitlement, will increase an understanding that may inform best practices in higher education. Use of the AE scale will allow researchers to identify an individual difference that predicts student incivility in higher education. The AE scale makes a unique contribution by explaining additional variance in students' academic behaviors, beyond that of previously published scales. The AE subscales of students' Externalized Responsibility and Entitled Expectations about their education capture stable individual differences relevant to the academic setting. The AE scale has potential uses, both in personality research as well as in the classroom, to identify and reeducate students who possess faulty expectations about college-level work.

References

Abouserie, R. (1994). Sources and level of stress in relation to locus of control and self-esteem in college students. *Educational Psychology, 14,* 323–330.

Amada, G. (1999). *Coping with misconduct in the college classroom: A practical model.* Asheville, NC: College Administration Publications.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bogart, L. M., Benotsch, E. G., & Pavlovic, J. D. (2004). Feeling superior but threatened: The relation of narcissism to social comparison. *Basic and Applied Social Psychology, 26,* 35–44.

Boice, R. (1996). *First-order principles for college teachers: Ten basic ways to improve the teaching process.* Bolton, MA: Anker Publishing.

Boudreau, C. A., & Kromrey, J. D. (1994). A longitudinal study of the *retention* and *academic performance* of participants in freshmen orientation course. *Journal of College Student Development, 35,* 444–449.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42,* 116–131.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48,* 306–307.

Campbell, W. K., Bonacci, A. M., Shelton, J., Exline, J. J., & Bushman, B. J. (2004). Psychological entitlement: Interpersonal consequences and validation of a self-report measure. *Journal of Personality Assessment, 83,* 29–45.

Cantwell, R. H., & Moore, P. J. (1996). The development of measures of individual differences in self-regulatory control and their relationship to academic performance. *Contemporary Educational Psychology, 21,* 500–517.

Carbone, E. (1998). *Teaching large classes: Tools and strategies.* Thousand Oaks, CA: Sage.

Carbone, E. (1999). Students behaving badly in large classes. *New Directions for Teaching and Learning, 77,* 35–43.

Cattell, R. B. (1978). *The scientific use of factor analysis.* New York: Plenum.

Constantinople, A., Cornelius, R., & Gray, J. (1988). The chilly climate: Fact or fiction? *Journal of Higher Education, 59,* 527–550.

Crawford, M., & MacLeod, M. (1990). Gender in the college classroom: An assessment of the "chilly climate" for women. *Sex Roles, 23,* 101–122.

Crocker, J., Luhtanen, R. K., Cooper, M. L., & Bouvrette, S. (2003). Contingencies of self-worth in college students: Theory and measurement. *Journal of Personality and Social Psychology, 85,* 894–908.

Crocker, J., Sommers, S. R., & Luhtanen, R. K. (2002). Hopes dashed and dreams fulfilled: Contingencies of self-worth and graduate school admissions. *Personality and Social Psychology Bulletin, 28,* 1275–1286.

Crosnoe, R., & Huston, A. C. (2007). Socioeconomic status, schooling, and the developmental trajectories of adolescents. *Developmental Psychology, 43,* 1097–1110.

Davis, G. H., & Mettee, D. R. (1971). Internal versus external control and magnitude of aggression toward self and others. *Psychological Reports, 29,* 403–411.

Diener, E., Lusk, R., DeFour, D., & Flax, R. (1980). Deindividuation: Effects of group size, density, number of observers, and group member similarity on self-consciousness and disinhibited behavior. *Journal of Personality and Social Psychology, 39,* 449–459.

Earl, W. R. (1987). Intrusive advising of freshmen in academic difficulty. *National Academic Advising Association Journal, 8,* 27–33.

Emmons, R. A. (1987). Narcissism: Theory and measurement. *Journal of Personality and Social Psychology, 52,* 11–17.

Exline, J. J., Baumeister, R. F., Bushman, B. J., Campbell, W. K., & Finkel, E. J. (2004). Too proud to let go: Narcissistic entitlement as a barrier to forgiveness. *Journal of Personality and Social Psychology, 87,* 894–912.

Feather, N. T. (1969). Attribution of responsibility and valence of success and failure in relation to initial confidence and task performance. *Journal of Personality and Social Psychology, 13,* 129–144.

Feather, N. T., & Simon, J. G. (1972). Causal attributions for success and

failure in relation to expectations of success based upon selective or manipulative control. *Journal of Personality, 39,* 527–541.

Feldman, R. S., Saletsky, R. D., Sullivan, J., & Theiss, A. (1983). Student locus of control and response to expectations about self and teacher. *Journal of Educational Psychology, 75,* 27–32.

Felsten, G., & Wilcox, K. (1992). Influences of stress, situation-specific mastery beliefs and satisfaction with social support on well-being and academic performance. *Psychological Reports, 70,* 219–303.

Flouri, E. (2006). Parental interest in children's education, children's self-esteem and locus of control, and later educational attainment: Twenty-six year follow-up of the 1970 British Birth Cohort. *British Journal of Educational Psychology, 76,* 41–55.

Glennen, R. E., Baxley, D. M., & Farren, P. J. (1985). Impact of intrusive advising on minority student retention. *College Student Journal, 19,* 335–338.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Graham, S., & Weiner, B. (1996). Theories and principles of motivation. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 63–84). New York: Macmillan.

Gump, S. E. (2006). Guess who's (not) coming to class: Student attitudes as indicators of attendance. *Educational Studies, 32,* 39–46.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory—Versions 4a and 54.* Berkeley: University of California, Institute of Personality and Social Research.

Joreskog, K. G., & Sorbom, D. G. (1996). *LISREL 8: User's reference guide.* Chicago: Scientific Software.

Kerr, S., Johnson, V. K., Gans, S. E., & Krumrine, J. (2004). Predicting adjustment during the transition to college: Alexithymia, perceived stress, and psychological symptoms. *Journal of College Student Development, 45,* 593–611.

Kim, J., & Mueller, C. (1978). *Factor analysis: Statistical methods and practical issues.* Beverly Hills, CA: Sage.

Kubarych, T. S., Deary, I. J., & Austin, E. J. (2004). The Narcissistic Personality Inventory: Factor structure in a non-clinical sample. *Personality and Individual Differences, 36,* 857–872.

Louie, T. A., & Tom, G. (2005). Timely completion of early class requirements: Effects of student and faculty gender. *Sex Roles, 52,* 245–250.

McGlynn, A. P. (2001). *Successful beginnings for college teaching: Engaging your students from the first day.* Madison, WI: Atwood Publishing.

Meyers, S. A. (2003). Strategies to prevent and reduce conflict in college classrooms. *College Teaching, 51,* 94–98.

Möller, J., & Köller, O. (2000). Spontaneous and reactive attributions following academic achievement. *Social Psychology of Education, 4,* 67–86.

Paulhus, D. L. (1983). Sphere-specific measures of perceived control. *Journal of Personality and Social Psychology, 44,* 1253–1265.

Perry, R. P., & Magnusson, J. (1989). Causal attributions and perceived performance: Consequences for college students' achievement and perceived control in different instructional conditions. *Journal of Educational Psychology, 81,* 164–172.

Perry, R. P., & Penner, K. S. (1990). Enhancing academic achievement in college students through attributional retraining and instruction. *Journal of Educational Psychology, 82,* 262–271.

Raskin, R., & Hall, C. S. (1979). A Narcissistic Personality Inventory. *Psychological Reports, 45,* 590.

Raskin, R., & Hall, C. S. (1981). The Narcissistic Personality Inventory: Alternate form reliability and further evidence of construct validity. *Journal of Personality Assessment, 45,* 159–162.

Rosenberg, M. (1989). *Society and the adolescent self-image* (Rev. ed.). Middletown, CT: Wesleyan University Press.

Rosenthal, S. A., Hooley, J. M., & Steshenko, Y. (2003, February). *Distinguishing grandiosity from self-esteem: Development of the State-Trait Grandiosity Scale.* Poster presented at the 4th Annual Meeting of the Society for Personality and Social Psychology, Los Angeles.

Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology.* New York: McGraw Hill.

Sanchez, R. J., Bauer, T. N., & Paronto, M. E. (2006). Peer-mentoring freshmen: Implications for satisfaction, commitment, and retention to graduation. *Academy of Management Learning & Education, 5,* 25–37.

Santiago-Rivera, A. L., & Bernstein, B. L. (1996). Affiliation, achievement and life events: Contributors to stress appraisals in college men and women. *Personality and Individual Differences, 21,* 411–419.

Schee, B. A. V. (2007). Adding insight to intrusive advising and its effectiveness with students on probation. *National Academic Advising Association Journal, 27,* 50–59.

Shell, D. F., & Husman, J. (2008). Control, motivation, affect, and strategic self-regulation in the college classroom: A multidimensional phenomenon. *Journal of Educational Psychology, 100,* 443–459.

Smalley, R. L., & Stake, J. E. (1996). Evaluating sources of ego-threatening feedback: Self-esteem and narcissism effects. *Journal of Research in Personality, 30,* 483–495.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180.

Stucke, T. S., & Sporer, S. L. (2002). When a grandiose self-image is threatened: Narcissism and self-concept clarity as predictors of negative emotions and aggression following ego-threat. *Journal of Personality, 70,* 509–532.

Stupnisky, R. H., Renaud, R. D., Perry, R. P., Ruthig, J. C., Haynes, T. L., & Clifton, R. A. (2007). Comparing self-esteem and perceived control as predictors of first-year students' academic achievement. *Social Psychology of Education, 10,* 303–330.

Tiberius, R. G., & Flak, E. (1999). Incivility in dyadic teaching and learning. *New Directions in Teaching and Learning, 77,* 3–12.

Ting, S. R., Grant, S., & Plenert, S. L. (2000). The Excellence-Commitment-and-Effective-Learning (EXCEL) group: An integrated approach for first-year college students' success. *Journal of College Student Development, 41,* 353–360.

Tom, G. (1998). Faculty and student perceptions of classroom etiquette. *Journal of College Student Development, 39,* 515–517.

Toothaker, L. E., & Miller, L. (1996). *Introductory statistics for the behavioral sciences* (2nd ed.). Pacific Grove, CA: Thomson.

Twenge, J. M., Zhang, L., & Im, C. (2004). It's beyond my control: A cross-temporal meta-analysis of increasing externality in locus of control, 1960–2002. *Personality and Social Psychology Review, 8,* 308–319.

Watson, D., & Clark, L. A. (1994). *Manual for the Positive and Negative Affect Schedule (Expanded Form).* Iowa City: University of Iowa.

Williams, C. B., & Vantress, F. E. (1969). Relation between internal–external control and aggression. *Journal of Psychology: Interdisciplinary and Applied, 71,* 59–61.

Wulff, D. H., Nyquist, J. D., & Abbott, R. D. (1987). Students' perceptions of large classes. In M. G. Weimer (Ed.), *Teaching large classes well* (pp. 17–30). San Francisco: Jossey-Bass.

Young, T. J. (1992). Locus of control and perceptions of human aggression. *Perceptual and Motor Skills, 74,* 1016–1018.

Received April 3, 2008
Revision received May 1, 2009
Accepted May 4, 2009 ■

Acknowledgments

The editor thanks the following principal reviewers who evaluated at least 4 manuscripts for *Journal of Educational Psychology* between January 1, 2008 and May 31, 2009.

Mary D. Ainley
Vincent Aleven

Andrew Biemiller
Eric S. Buhs

Clark Chinn
Lyn Corno
Scott A. Crossley

Sidney K. D'Mello

Heather A. Davis
David K. Dickinson
Amanda M. Durik

Jeffrey A. Greene
Frederic Guay
John Guthrie

John A. C. Hattie
Jan N. Hughes

Tanner Jackson
Martin H. Jones

Carol Anne Kardash
Allison J. Kelaher Young

Willy Lens
Nonie K. Lesaux

Beth Meisinger

Tamara Murdock

Kristie J. Newton

Tenaha O'Reilly

Lindsey Richland

Dale H. Schunk

The editor wishes to thank those who reviewed manuscripts between January 1, 2008 and May 31, 2009.

Karen Ablard Mickenberg
Steve Alessi
Joyce Alexander
Richard Allington
Janice F. Almasi
Jesus Alonso-Tapia
Lynley H. Anderman
Richard C. Anderson
Alison Arrow
Jane Ashby
Mark Ashcraft

Linda Baker
Galen Baril
Ken Barron
Ann A. Battle
Margaret Beebe-Frankenberger
Avi (Talia) Ben-Zeev
Alpana Bhattacharya
Kathy S. Binder
Linda Bol
Julie Booth
Donald L. Bolger
Harry Brenton
Jere Brophy
B. Bradford Brown
Stephen Burgess
Roger Bruning
Michelle Buehl
Kirsten R. Butcher
James L. Byo
Brian Byrne
James Byrnes
Soo-yong Byun

Michelene T. H. Chi
Tim J. Cleary
Catherine A. Clement
James L. Collins
Roberto Colom
Susan Colmar
Pierre Cormier
Jennifer Corpus
Leslie J. Couse
Rhonda G. Craven
Robert Cosnoe
Timothy Curby
Reagan Curtis

Bridget Dalton

Jodi Davenport
Mark Davis
Siegfried Dewitte
Janice Dole

Anastasia D. Elder
Cynthia A. Erdley
Mary Ann Evans
E. Margaret Evans
Howard T. Everson

Gary Feng
Daniel Flannery
Jack Fletcher
Anne Foegen
Donna Y. Ford
David Francis
Karin S. Frey
Douglas Fuchs
Adrian F. Furnham

Linda B. Gambrell
Russell Gersten
Richard C. Gilman
Paul Ginns
Jane Ginsborg
Usha Goswami
Adele E. Gottried
Sandra Graham
Kathy E. Green
Thomas D. Griffin
Wendy Grolnick

Zach Hambrick
Paul E. Hand
Laurie B. Hanich
James Hartley
Virginia Smith Harvey
Mary Hegarty
Robin K. Henson
Suzannye Hidi
Heather Hill
Cindy Hmelo-Silver
Barbara K. Hofer
Don Hossler
Cynthia Hudley
Janet S. Hyde

John Jabagchourian

William Jeynes
Jenelle Job
Laura Justice
Jaana Juvonen

Slava Kalyuga
Stuart A. Karabenick
James C. Kaufman
Ronald T. Kellogg
Panayiota Kendeou
Kenneth Kiewra
Walter Kintsch
Eileen Kitsch
Paul A. Kirschner
Anastasia Kitsantas
Ingo Kollar
Andreas Krapp
Evelyn Kroesbergen
Li-Jen Kuo
Christopher A. Kurby
Beth Kurtz-Costes

Kimberly A. Lawless
A. Michele Lease
Richard Lehrer
Lauren Liang
Phil D. Liu
David F. Lohman
Robert F. Lorch
Jon P. Lorence
Marsha Lovett
Shulan Lu
Heikki J. Lyytinen

Xin Ma
Joseph P. Magliano
Roxana Marachi
Andrew J. Mashburn
Deborah McCutchen
Nicole McNeil
Fred J. Medway
Judith Meece
Bonnie J. F. Meyer
Kevin F. Miller
David Moore
Daniel Moos
Judit N. Moschkovich
Chris Mueller
Krista R. Muis

Kou Murayama

John C. Nesbit
Florrie Ng
Timothy J. Nokes
E. Michael Nussbaum

Susan O'Neill
Morris Okrun
Richard Olson
Jason Osborne
Yasuhiro Ozuru

Rene Parmar
Reinhard Pekrun
Beth Phillips
Therese D. Pigott
Paul Poteat
Patrick Proctor

Darshanand Ramdass
Keith Rayner
Robert Reid
Alenxander Renkl
K. Ann Renninger
Matthew Reynolds
Cara Richards
John T. E. Richardson
Katherine Robinson
Phil Rodkin
Jeremy Roschelle
Rod Roscoe
Julie Rosenthal
Kathleen Moritz Rudasill
Nikol Rummel
Allison M. Ryan

John Sabatini
Ladislao Salmeron
Wolfgang Schneider
Paula Schwanenflugel
David Schwartz
Dan Schwartz
Karl Schweizer
Timothy Shanahan
Sungok Serena Shim
Hua Shu
Georgios D. Sideridis
Robert S. Siegler

Greg Simpson
EllenSkinner
Jessi Smith
Bret P. Smith
Lisa M. Soederberg Miller
Marcantonio M. Spada
Deborah Speece
Rayne A. Sperling
Jon R. Star
Hillary H. Steiner
Donald M. Stenhoff
Robert J. Sternberg
Steven Stroessner
H. Lee Swanson

Raphael Taffy
Sigmund Tobias
Wendy Troop-Gordon
Michelle Trotman Scott
Mary P. Truxaw
Megan Tschannen-Moran
Julianne C. Turner

Giovanni Valiante
Robert J. Vallerand
Brandon Vaughn
Eduardo Vidal-Abarca
Joan M. T. Walker
Claire Ellen Weinstein
Mary Jane White
David M. Williams
Christopher R. Wolfe

Karen M. Zabrucky
Akane Zusho

Instructions to Authors
Journal of Educational Psychology
www.apa.org/journals/edu

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Manuscript preparation. Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (6th ed.). Manuscripts may be copyedited for bias-free language (see pp. 70–77 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see www.apa.org/journals. **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 250 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Haag, L., & Stern, E. (2003). In search of the benefits of learning Latin. *Journal of Educational Psychology, 95,* 174–178.

Johnson, D. W., & Johnson, R. T. (1990). Cooperative learning and achievement. In S. Sharon (Ed.), *Cooperative learning: Theory and research* (pp. 173–202). New York: Praeger.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual.* Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see p. 34 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied as Tiff, EPS, or PowerPoint. High-quality printouts or glossies are needed for *all* figures. The minimum line weight for line art is 0.5 point for optimal printing. When possible, please place symbol legends below the figure image instead of to the side. Original color figures can be printed in color at the editor's and publisher's discretion provided the author agrees to pay $255 for one figure, $425 for two figures, $575 for three figures, $675 for four figures, and $55 for each additional figure.

Publication policies. APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at www.apa.org/journals. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

Masked review policy. The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., "in our previous work, Johnson et al., 1998 reported that . . .." Instead, references to the authors' work should be in third person, e.g., "Johnson et al. (1998) reported that . . .." The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at www.apa.org/ethics/ or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

Permissions. Authors of accepted papers are required to obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including, for example, test materials or portions thereof and photographs of people.

Supplemental materials. APA can now place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see www.apa.org/journals/authors/suppmaterial.html for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

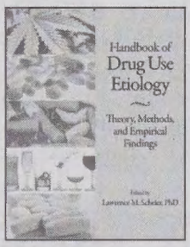
Submission. Authors should submit their manuscripts electronically via the Manuscript Submission Portal at www.apa.org/journals/edu (follow the link "Submit Manuscripts Electronically"). A checklist for manuscript submission, including guidelines for preparing the electronic file, can be found at www.apa.org/journals/. Correspondence regarding manuscripts should be sent to the Editor, Art Graesser, University of Memphis, Journal of Educational Psychology, 202 Psychology Building, Memphis, TN 38152-3230. In addition to addresses and phone numbers, authors should supply e-mail addresses, as most communications will be by e-mail. Fax numbers, if available, should also be provided for potential use by the editorial office and later by the production office. Authors should keep a copy of the manuscript to guard against loss. E-mail correspondence may be addressed to jedgar@memphis.edu.

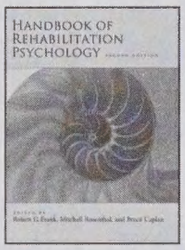
Preparing files for production. If your manuscript is accepted for publication, please follow the guidelines for file formats and naming provided at www.apa.org/journals/authors/preparing_efiles.html. If your manuscript was mask reviewed, please ensure that the final version for production includes a byline and full author note for typesetting.

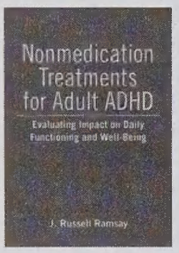
NEW RELEASES

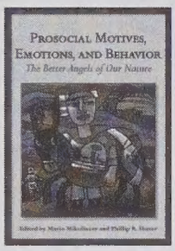
from the American Psychological Association

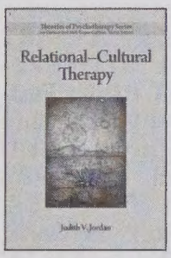
AMERICAN PSYCHOLOGICAL ASSOCIATION

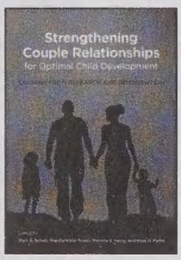


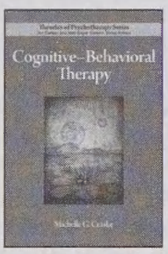


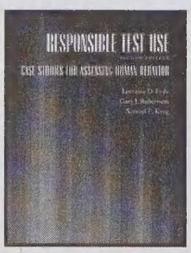


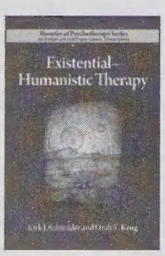


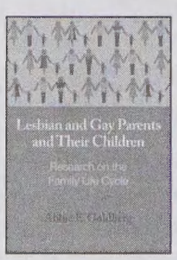












Handbook of Drug Use Etiology

Theory, Methods, and Empirical Findings

Edited by Lawrence M. Scheier

2010. 784 pages. Hardcover.
ISBN 978-1-4338-0446-5; Item # 4311501
List: $129.95; APA Member/Affiliate: $69.95

Nonmedication Treatments for Adult ADHD

Evaluating Impact on Daily Functioning and Well-Being

J. Russell Ramsay

2010. 232 pages. Hardcover.
ISBN 978-1-4338-0564-6; Item # 4317203
List: $69.95; APA Member/Affiliate: $49.95

Relational–Cultural Therapy

Judith V. Jordan

2010. 160 pages. Paperback.
ISBN 978-1-4338-0463-2; Item # 4317194
List: $24.95; APA Member/Affiliate: $24.95

Cognitive–Behavioral Therapy

Michelle G. Craske

2010. 192 pages. Paperback.
ISBN 978-1-4338-0548-6; Item # 4317199
List: $24.95; APA Member/Affiliate: $24.95

Feminist Therapy

Laura S. Brown

2010. 160 pages. Paperback.
ISBN 978-1-4338-0461-8; Item # 4317192
List: $24.95; APA Member/Affiliate: $24.95

Existential–Humanistic Therapy

Kirk J. Schneider and Orah T. Krug

2010. 176 pages. Paperback.
ISBN 978-1-4338-0462-5; Item # 4317193
List: $24.95; APA Member/Affiliate: $24.95

Lesbian and Gay Parents and Their Children

Research on the Family Life Cycle

Abbie E. Goldberg

2010. 232 pages. Hardcover.
ISBN 978-1-4338-0536-3; Item # 4318061
List: $69.95; APA Member/Affiliate: $49.95

Handbook of Rehabilitation Psychology

SECOND EDITION

Edited by Robert G. Frank, Mitchell Rosenthal, and Bruce Caplan

2010. 584 pages. Hardcover.
ISBN 978-1-4338-0444-1; Item # 4311500
List: $99.95; APA Member/Affiliate: $59.95

Prosocial Motives, Emotions, and Behavior

The Better Angels of Our Nature

Edited by Mario Mikulincer and Phillip R. Shaver

2010. 448 pages. Hardcover.
ISBN 978-1-4338-0546-2; Item # 4318062
List: $89.95; APA Member/Affiliate: $59.95

Strengthening Couple Relationships for Optimal Child Development

Lessons From Research and Intervention

Edited by Marc S. Schulz, Marsha Kline Pruett, Patricia K. Kerig, and Ross D. Parke

2010. 272 pages. Hardcover.
ISBN 978-1-4338-0547-9; Item # 4318064
List: $79.95; APA Member/Affiliate: $49.95

Helping Families and Communities Recover From Disaster

Lessons Learned From Hurricane Katrina and Its Aftermath

Edited by Ryan P. Kilmer, Virginia Gil-Rivas, Richard G. Tedeschi, and Lawrence G. Calhoun

2010. 328 pages. Hardcover.
ISBN 978-1-4338-0544-8; Item # 4316114
List: $69.95; APA Member/Affiliate: $49.95

Responsible Test Use

Case Studies for Assessing Human Behavior

SECOND EDITION

Lorraine D. Eyde, Gary J. Robertson, and Samuel E. Krug

2010. 232 pages. Paperback.
ISBN 978-1-4338-0556-1; Item # 4311013
List: $39.95; APA Member/Affiliate: $29.95

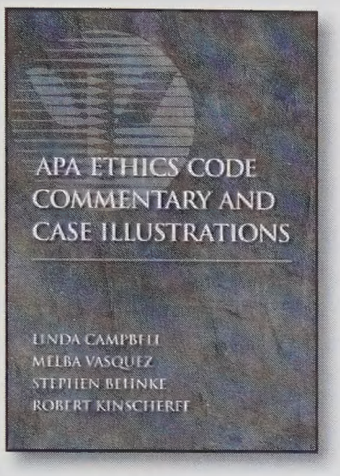
AD0690

To Order: 800-374-2721 • www.apa.org/books

NEW RELEASES

from the American Psychological Association

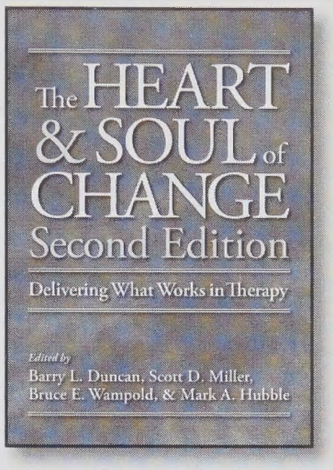
AMERICAN PSYCHOLOGICAL ASSOCIATION



APA Ethics Code Commentary and Case Illustrations

Linda Campbell, Melba Vasquez, Stephen Behnke, and Robert Kinscherff

2010. 392 pages. Hardcover.
ISBN 978-1-4338-0693-3
Item # 4312015
List: $69.95
APA Member/Affiliate: $49.95



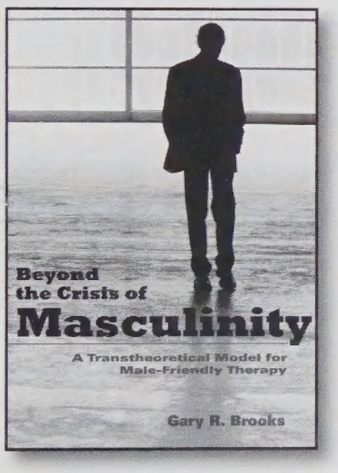
The Heart and Soul of Change

SECOND EDITION

Delivering What Works in Therapy

Edited by Barry L. Duncan, Scott D. Miller, Bruce E. Wampold, and Mark A. Hubble

2010. 464 pages. Hardcover.
ISBN 978-1-4338-0709-1
Item # 4317206
List: $59.95
APA Member/Affiliate: $49.95

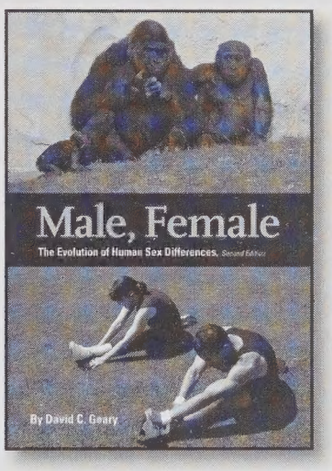


Beyond the Crisis of Masculinity

A Transtheoretical Model for Male-Friendly Therapy

Gary R. Brooks

2010. 232 pages. Hardcover.
ISBN 978-1-4338-0716-9
Item # 4317207
List: $69.95
APA Member/Affiliate: $49.95



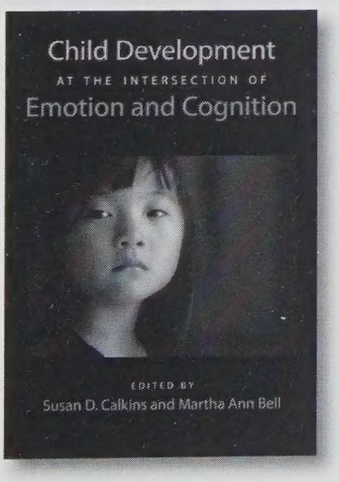
Male, Female

The Evolution of Human Sex Differences

SECOND EDITION

David C. Geary

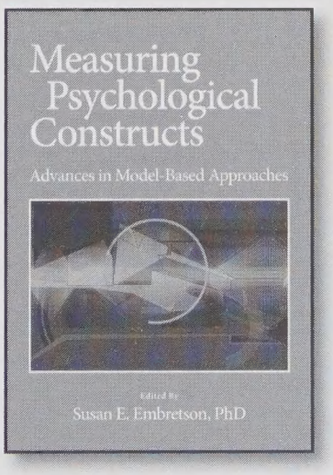
2010. 568 pages. Hardcover.
ISBN 978-1-4338-0682-7
Item # 4318066
List: $69.95
APA Member/Affiliate: $49.95



Child Development at the Intersection of Emotion and Cognition

Edited by Susan D. Calkins and Martha Ann Bell

2010. 264 pages. Hardcover.
ISBN 978-1-4338-0686-5
Item # 4318067
List: $79.95
APA Member/Affiliate: $49.95



Measuring Psychological Constructs

Advances in Model-Based Approaches

Edited by Susan E. Embretson

2010. 312 pages. Hardcover.
ISBN 978-1-4338-0691-9
Item # 4318069
List: $79.95
APA Member/Affiliate: $49.95

AD0698

To Order: 800-374-2721 • www.apa.org/books